AICT 2014

The Tenth Advanced International Conference on Telecommunications

ISBN: 978-1-61208-360-5

July 20 - 24, 2014

Paris, France

AICT 2014 Editors

Michael Logothetis, University of Patras, Greece
Michael Massoth, University of Applied Sciences - Darmstadt, Germany
Tulin Atmaca, Telecom SudParis, France
Eugen Borcoci, University Politehncia Bucharest, Romania

AICT 2014

Foreword

The Tenth Advanced International Conference on Telecommunications (AICT 2014), held between July 20-24, 2014, in Paris, France, covered a variety of challenging telecommunication topics ranging from background fields like signals, traffic, coding, communication basics up to large communication systems and networks, fixed, mobile and integrated, etc. Applications, services, system and network management issues also received significant attention.

The spectrum of 21st Century telecommunications is marked by the arrival of new business models, new platforms, new architectures and new customer profiles. Next generation networks, IP multimedia systems, IPTV, and converging network and services are new telecommunications paradigms. Technology achievements in terms of co-existence of IPv4 and IPv6, multiple access technologies, IP-MPLS network design driven methods, multicast and high speed require innovative approaches to design and develop large scale telecommunications networks.

Mobile and wireless communications add profit to large spectrum of technologies and services. We witness the evolution 2G, 2.5G, 3G and beyond, personal communications, cellular and ad hoc networks, as well as multimedia communications.

Web Services add a new dimension to telecommunications, where aspects of speed, security, trust, performance, resilience, and robustness are particularly salient. This requires new service delivery platforms, intelligent network theory, new telecommunications software tools, new communications protocols and standards.

We are witnessing many technological paradigm shifts imposed by the complexity induced by the notions of fully shared resources, cooperative work, and resource availability. P2P, GRID, Clusters, Web Services, Delay Tolerant Networks, Service/Resource identification and localization illustrate aspects where some components and/or services expose features that are neither stable nor fully guaranteed. Examples of technologies exposing similar behavior are WiFi, WiMax, WideBand, UWB, ZigBee, MBWA and others.

Management aspects related to autonomic and adaptive management includes the entire arsenal of self-ilities. Autonomic Computing, On-Demand Networks and Utility Computing together with Adaptive Management and Self-Management Applications collocating with classical networks management represent other categories of behavior dealing with the paradigm of partial and intermittent resources.

We take here the opportunity to warmly thank all the members of the AICT 2014 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to AICT 2014. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the AICT 2014 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that AICT 2014 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of telecommunications.

We are convinced that the participants found the event useful and communications very open. We hope that Paris, France, provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

AICT 2014 Chairs:

AICT Advisory Committee
Tulin Atmaca, Telecom SudParis, France
Eugen Borcoci, University Politehncia Bucharest, Romania
Michael D. Logothetis, University of Patras, Greece
Go Hasegawa, Osaka University, Japan
Reijo Savola, VTT Technical Research Centre of Finland - Oulu, Finland
Michael Massoth, University of Applied Sciences - Darmstadt, Germany
Mariusz Glabowski, Poznan University of Technology, Poland
Djafar K. Mynbaev, New York City College of Technology - Brooklyn, USA
Dragana Krstic, Faculty of Electronic Engineering, University of Nis, Serbia
Ruediger Gad, University of Applied Sciences Frankfurt am Main, Germany
Erchin Serpedin, Texas A&M University, USA
Mohammed Al-Olofi, Duisburg-Essen University, Germany

AICT Industry/Research Chairs
Andres Arjona, Nokia Siemens Networks, Japan
Michael Atighetchi, Raytheon BBN Technologies-Cambridge, USA
Kazuya Tsukamoto, Kyushu Institute of Technology-Fukuoka, Japan
Guillaume Valadon, French Network and Information Security Agency, France
Sergei Semenov, Broadcom, Finland
Abheek Saha, Hughes Systique Corporation, USA
John Vardakas, Iquadrat Barcelona, Spain
Sladjana Zoric, Deutsche Telekom AG, Bonn, Germany
Christophe Feltus, Public Research Center Henri Tudor, Luxembourg
Hussein Kdouh, IETR, France
Yasunori Iwanami, Nagoya Institute of Technology, Japan

AICT Publicity Chair
Ustijana Rechkoska Shikoska, University for Information Science & Technology "St. Paul The Apostle" - Ohrid, Republic of Macedonia

AICT 2014

COMMITTEE

AICT Advisory Committee

Tulin Atmaca, Telecom SudParis, France
Eugen Borcoci, University Politehncia Bucharest, Romania
Michael D. Logothetis, University of Patras, Greece
Go Hasegawa, Osaka University, Japan
Reijo Savola, VTT Technical Research Centre of Finland - Oulu, Finland
Michael Massoth, University of Applied Sciences - Darmstadt, Germany
Mariusz Glabowski, Poznan University of Technology, Poland
Djafar K. Mynbaev, New York City College of Technology - Brooklyn, USA
Dragana Krstic, Faculty of Electronic Engineering, University of Nis, Serbia
Ruediger Gad, University of Applied Sciences Frankfurt am Main, Germany
Erchin Serpedin, Texas A&M University, USA
Mohammed Al-Olofi, Duisburg-Essen University, Germany

AICT Industry/Research Chairs

Andres Arjona, Nokia Siemens Networks, Japan
Michael Atighetchi, Raytheon BBN Technologies-Cambridge, USA
Kazuya Tsukamoto, Kyushu Institute of Technology-Fukuoka, Japan
Guillaume Valadon, French Network and Information Security Agency, France
Sergei Semenov, Broadcom, Finland
Abheek Saha, Hughes Systique Corporation, USA
John Vardakas, Iquadrat Barcelona, Spain
Sladjana Zoric, Deutsche Telekom AG, Bonn, Germany
Christophe Feltus, Public Research Center Henri Tudor, Luxembourg
Hussein Kdouh, IETR, France
Yasunori Iwanami, Nagoya Institute of Technology, Japan

AICT Publicity Chair

Ustijana Rechkoska Shikoska, University for Information Science & Technology "St. Paul The Apostle" - Ohrid, Republic of Macedonia

AICT 2014 Technical Program Committee

Fatma Abdelkefi, High School of Communications of Tunis - SUPCOM, Tunisia
Sachin Kumar Agrawal, Samsung Electronics, India
Mahdi Aiash, Middlesex University - London, UK
Anwer Al-Dulaimi, Brunel University - Middlesex, UK
Sabapathy Ananthi, University of Madras, India
Josephina Antoniou, University of Central Lancashire, Cyprus
Pedro A. Aranda Gutiérrez, University of Paderborn, Germany

Alexandru Martian, Politehnica University of Bucharest, Romania
Michael Massoth, Hochschule Darmstadt, Germany
Martin May, Techniclor, France
Natarajan Meghanathan, Jackson State University, USA
Jean-Marc Menaud, École des Mines de Nantes / INRIA, LINA, France
Lynda Mokdad, Université Paris-Est-Créteil, France
Miklós Molnár, LIRMM/University of Montpellier II, France
Philip Morrow, University of Ulster-Coleraine, Northern Ireland, UK
Ioannis Moscholios, University of Peloponnese, Greece
Petr Münster, Brno University of Technology, Czech Republic
Juan Pedro Muñoz-Gea, Universidad Politécnica de Cartagena, Spain
Masayuki Murata, Osaka University, Japan
Djafar K. Mynbaev, New York City College of Technology - Brooklyn, USA
David Naccache, Université Paris II/Ecole normale supérieure, France
Amor Nafkha, SUPELEC, France
Antonio Navarro Martín, Universidad Complutense de Madrid, Spain
Nikolai Nefedov, ETH Zürich, Switzerland
Petros Nicopolitidis, Aristotle University of Thessaloniki, Greece
Serban Obreja, University “Politehnica” Bucharest, Romania
Niyazi Odabasioglu, Istanbul University, Turkey
Masaya Okada, Shizuoka University, Japan
Minoru Okada, Nara Institute of Science and Technology, Japan
Sema Oktug, Istanbul Technical University, Turkey
Cristina Oprea, Politehnica University of Bucharest, Romania
Ali Ozen, Nuh Naci Yazgan University, Turkey
Constantin Paleologu, University Politehnica of Bucharest, Romania
Jari Palomäki, Tampere University of Technology - Pori, Finland
Andreas Papazois, RACTI & CEID / University of Patras, Greece
Woogoo Park, ETRI, South Korea
Cathryn Peoples, University of Ulster, UK
Fernando Pereñíguez García, Universidad Católica San Antonio Murcia, Spain
Jordi Pérez Romero, Universitat Politecnica de Catalunya (UPC) - Barcelona, Spain
Maciej Piechowiak, Kazimierz Wielki University - Bydgoszcz, Poland
Michael Piotrowski, University of Zurich, Switzerland
Adrian Popescu, Blekinge Institute of Technology - Karlskrona, Sweden
Neeli R. Prasad, Aalborg University, Denmark
Emanuel Puschita, Technical University of Cluj-Napoca, Romania
Dusan Radovic, TES Electronic Solutions GmbH - Stuttgart, Germany
Adib Rastegarnia, University of Tehran, Iran
Ustijana Rechkoska Shikoska, University for Information Science & Technology "St. Paul the Apostle" - Ohrid, Republic of Macedonia
Yenumula Reddy, Grambling State University, USA
Eric Renault, Telecom SudParis, France
Lorayne Robertson, University of Ontario Institute of Technology, Canada
Pawel Rózycki, University of IT and Management (UITM), Poland
Danguole Rutkauskiene, Kaunas University of Technology, Lithuania
Abheek Saha, Hughes Systique Corporation, USA
Ramiro Sámano Robles, Instituto de Telcomunicaçoes, Portugal

Sladjana Zoric, Deutsche Telekom AG, Bonn, Germany
Piotr Zwierzykowski , Poznan University of Technology, Poland

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Biomedical Applications of Intensity and Curvature Measures: The Case of Magnetic Resonance Imaging of the Human Brain

Carlo Ciulla, Ustijana Reckoska Shikoska, Dijana Capeska Bogatinoska
University for Information Science & Technology
“St. Paul the Apostle”
Ohrid, Macedonia
e-mail: carlo.ciulla@uist.edu.mk, cxc2728@njit.edu,
ustijana@gmail.com, dijana.c.bogatinoska@uist.edu.mk

Filip A. Risteski, Dimitar Veljanovski
Skopje City General Hospital
Skopje, Macedonia
e-mail: risteskifilip@bolnica.org.mk,
dveljanovski@bolnica.org.mk

Abstract—**This paper intends to present a novel approach to the extraction of additional and/or complementary biomedical information from the Magnetic Resonance Imaging (MRI) of the human brain. The extraction of the biomedical information is conducted through three mathematical engineering tools called Classic-Curvature, Intensity-Curvature Functional and Intensity-Curvature Measure, which are calculated through a model function fitted to the MRI data. The mathematical engineering tools require that the model function benefits of the property of second-order differentiability. The Classic-Curvature, the Intensity-Curvature Functional and the Intensity-Curvature Measure descend from the unifying theory and the unified theory originally conceived for the improvement of the interpolation error. The advantage provided through the methodological approach is that an immense number of possible Classic-Curvature, Intensity-Curvature Functional and Intensity-Curvature Measure images can be derived through re-sampling at the intra-pixel coordinate, and this fact provides the possibility to choose images which give the best result in diagnostic practice. The biomedical information might be used in telemedicine.**

Keywords-Model Function; Classic-Curvature; Intensity-Curvature Functional; Signal-Image; Magnetic Resonance Imaging (MRI); Human Brain.

I. INTRODUCTION

The introduction section will describe the proposal, it will also describe why the theoretical basis of the present work differs from the state of the art and also it will outline the contribution of this paper.

A. *Description of the Approach*

Let us define the grid node as the location where sampling occurs in either one dimension (1D), two dimensions (2D), or three dimensions (3D). In a sequel of digital samples, in either 1D, 2D or 3D, let a given intra-node location be called the re-sampling location. The problem statement is given hereto: the calculation of three continuous math formulae from a discontinuous domain created by a sequel of digital samples. The requirement of the solution to the problem is that a model function, which embeds the property of second-order differentiability [1], needs to be fitted to the discontinuous domain.

The solution to the herein stated problem consists in the calculation of the Classic-Curvature at the re-sampling location [1]. Specifically, given an image and fitting the model function to the image, it is possible to calculate the Classic-Curvature through the summation of all of the partial second order derivatives of the Hessian [1] of the model function [1]. The partial second order derivatives are calculated at the re-sampling location. The re-sampling location is the intra-pixel coordinate where the signal-image is calculated through the model function.

The calculation of the Classic-Curvature makes it possible also to calculate the Intensity-Curvature Functional [2][3] at the re-sampling location as follows. The ratio between two terms: (i) the integral of the product between the signal intensity and the Classic-Curvature both of them calculated at the grid node; and (ii) the integral of the product between the signal intensity and the Classic-Curvature both of them calculated at the re-sampling location. The calculation of the Intensity-Curvature Measure has been introduced in [4].

B. *Comparison with other Solutions*

The literature shows a widespread use of approximations of the curvature of the signal-image through compact finite differences, and/or gradients and/or the Sobel operator [5] for the calculation of the first order derivative of the signal-image see, for instance, the work reported in [5]-[8]. The necessity of having a rigorous method, which is based on calculus, in order to quantify the curvature of the signal-image, makes the present paper different and unique in the field of biomedical signal-image processing. In fact, in this work, the calculation of the Classic-Curvature is made through all of the second-order partial derivatives of the Hessian of the model polynomial function fitted to the MRI data. The advantage is that of summing up all of the partial second order derivatives of the Hessian of the model function fitted to the image. By doing so, the covariates partial derivatives are included in the calculation of the Classic-Curvature, the Intensity-Curvature Functional and the Intensity-Curvature Measure.

C. *Applicability to the Life Sciences*

The proposal presented in this paper has the potential to contribute to life sciences because of the three novel images: Classic-Curvature, Intensity-Curvature Functional (see

Figure 1) and Intensity-Curvature Measure (see Figures 9 and 10), which embeds biomedical information content having diagnostic value. For example, in human brain imaging, for what pertains to: (i) the demarcation of anatomical structures, (ii) highlighting of the difference between gray and white matter; both in normal and in pathological biomedical images, and (iii) the extraction of additional and/or complementary information from pathological MRI. The connection between the research here presented and the Information Communication Technologies is in the field of Telemedicine. Figure 1 shows the Original MRI in (a), which is provided by the courtesy of OASIS database [9]-[14][15]. Figure 1b shows the Classic-Curvature of the MRI seen in (a). Figure 1c shows the Intensity-Curvature Functional of the MRI seen in (a). The image in Figure 1b is calculated with the two-dimensional Lagrange polynomial [3] when re-sampling of the misplacement of 0.1mm along both of x and y axis, whereas the image in Figure 1c is calculated with the bivariate linear function [16], when re-sampling of the misplacement of 0.01mm along the x axis and 0.01mm along the y axis. In Figure 1, the image in (c) shows a third dimension perpendicular to the imaging plane along with the difference between gray and white matter. Both of the images in (b) and (c) are contrast-brightness enhanced.

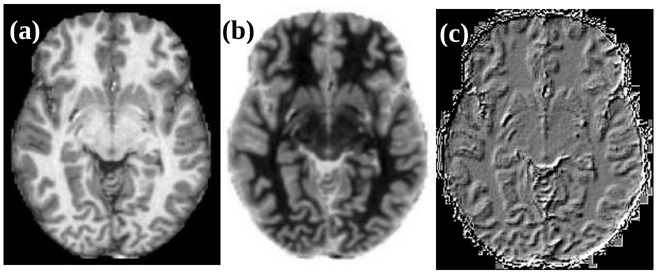


Figure 1. The image in (a) shows the original MRI and comprises of a 205x246 pixels matrix with 1.00mm x 1.00mm pixel size; (b) shows the Classic-Curvature of (a), which marks a clear difference between gray and white matter of the human brain and (c) shows the Intensity-Curvature Functional of (a).

In this paper, emphasis is given to four model functions, specifically: (i) the bivariate quadratic B-Spline polynomial [3], (ii) the bivariate cubic Lagrange polynomial [3], (iii) the one-dimensional Sinc function [2], and (iv) the bivariate linear function [16]. It is evident that the aforementioned four model functions are capable, through the application of the Classic-Curvature, the Intensity-Curvature Functional and the Intensity-Curvature Measure, to extract information from the MRI images, which is not readily observable into the original images.

Section II will focus on the capability of the Classic-Curvature, the Intensity-Curvature Functional and the Intensity-Curvature Measure to perform feature extraction from the original image. In Section III, the practical implications of this work will be addressed placing the emphasis on the methodological approach and also on the value added to the original MRI through the use of the three mathematical engineering tools used in this piece of research. Finally, Section IV concludes the paper.

II. RESULTS

This section presents qualitative results obtained through fitting to the MRI data of the human brain: (i) the bivariate quadratic B-Spline, (ii) the cubic Lagrange polynomial, and also (iii) the one-dimensional Sinc interpolation function [2]. Figure 2 shows two of the MRI images employed in this piece of research, which are referred here as to be the original MRI. Some of the Classic-Curvature and the Intensity-Curvature Functional reported in this section have been calculated on the basis of the images shown in Figure 2. The MRI image shown in Figure 2a is provided by the courtesy of Casa di Cure Triolo - Zancla, Palermo – Italy [3]. The MRI image shown in Figure 2b is provided by the courtesy of the OASIS database [15]. In Figure 2, the image in (a) has been scaled to enhance the visual appearance of the picture.

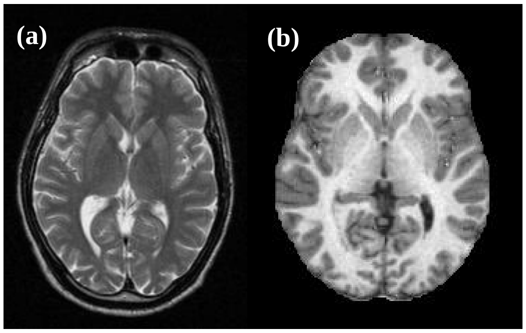


Figure 2. Original Magnetic Resonance Imaging data: (a) the image is made of a 176 x 234 pixels matrix with pixel size of 1mm x 1mm; (b) the image is made of a 176 x 208 pixels matrix with pixel size of 1mm x 1mm.

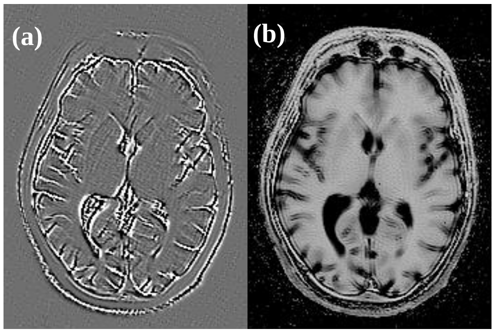


Figure 3. The image in (a) shows the Classic-Curvature and the image in (b) shows the Intensity-Curvature Functional. The brain structures highlighted in (a) are those of the sulci and the brain ventricles (see white contours). In (b) the emphasis is still on the sulci of the human brain and the depth is highlighted.

Figure 3 shows the Classic-Curvature image in (a) and the Intensity-Curvature Functional image in (b). Specifically, since it is the objective of this piece of research to assess the capability of two of the mathematical engineering tools to provide complementary information through feature extraction from the original MRI, the reader should compare the appearance of the Classic-Curvature and the Intensity-Curvature Functional images with the original MRIs shown in Figure 2. Both of the images in Figure 3 were obtained when fitting to the signal-data the bivariate quadratic B-Spline re-sampling of 0.01mm along the x direction and 0.01mm along the y direction. Both of the images in Figure 3 are contrast-brightness enhanced. Figure 4 shows two Intensity-Curvature Functional images obtained when fitting the bivariate quadratic B-Spline to the human brain data and when re-sampling was performed at the misplacement (x, y) ≡ (0.01mm, 0.001mm) with the value of the 'a' constant parameter set to 7 in both of (a) and (b). The difference observable between (a) and (b) is attributable to the pre-processing step, which standardizes (see (a)), and scales (see (b)) the pixel intensity respectively.

In Figure 4, the image in (a) shows a neat distinction between the gray and the white matter of the human brain, whereas the image in (b) shows the same distinction however with a third dimension visible in the direction perpendicular to the image plane. Both of the images are contrast-brightness enhanced.

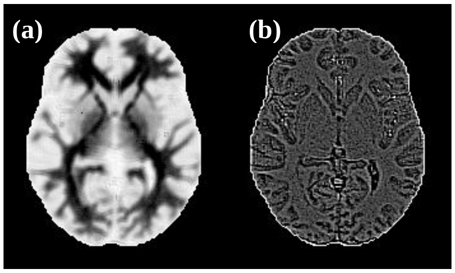


Figure 4. The images in (a) and (b) are both Intensity-Curvature Functional of the original MRI shown in Figure 2b and they were obtained when fitting the bivariate quadratic B-Spline to the signal data.

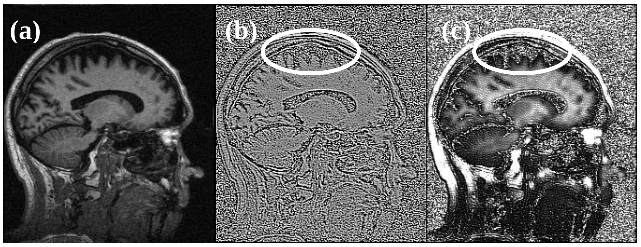


Figure 5. The image in (a) is the original MRI and the images in (b) and (c) are Intensity-Curvature Functional with visible and well demarcated brain anatomical structures.

Figure 5 shows two Intensity-Curvature Functional images in (b) and in (c) obtained from the original MRI shown in (a). The original MRI is provided by the courtesy of the OASIS database and was collected on a subject classified positive to the Clinical Dementia Rating (CDR) [10][12]. In Figure 5, the pixels matrix size is 256 x 256 with pixel size 1mm x 1mm. The Intensity-Curvature Functional images were obtained when fitting the bivariate quadratic B-Spline to the signal-image, specifically when using the 'a' constant parameter set to 3.54 (b) and -3.54 (c). The misplacement used for re-sampling is (x, y) ≡ (0.01mm, 0.01mm) in both of (b) and (c). What is remarkable in Figure 5 is the fact that the shrinkage of the human cortex, which is well visible in (a), is also visible in (b) and (c) where the value of the Intensity-Curvature Functional is comparable to the noise level of the rest of the image (see inside the white ellipses).

In Figure 5, the human cortex is distinguishable in both images (b) and (c). The images were cropped to highlight the regions of interest and they are contrast-brightness enhanced.

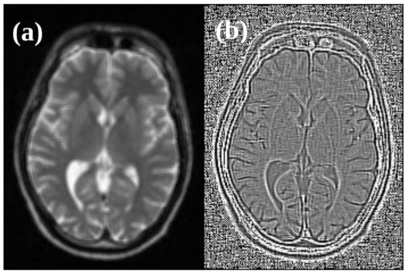


Figure 6. The image in (a) shows the Classic-Curvature and the image in (b) shows the Intensity-Curvature Functional. The two images were obtained when re-sampling with the bivariate cubic Lagrange polynomial with a misplacement of (x, y) ≡ (0.01mm, 0.01mm) in (a) and a misplacement of (x, y) ≡ (0.95mm, 0.95mm) in (b).

Figure 6 shows results obtained when fitting the bivariate cubic Lagrange polynomial to the brain image data. While the Classic-Curvature demonstrates faithful reproduction of the original MRI therefore showing all of the human brain features, the Intensity-Curvature Functional places the emphasis on the small features of the human brain such as the sulci showing well demarcated anatomy. The same can be said for the brain ventricles. In other words, the image depicted in (b) performs feature extraction from the image seen in (a), therefore showing details that are not readily seen in (a), neither in the original MRI. A similar behavior of both the Classic-Curvature and the Intensity-Curvature Functional was already observed in Figure 3. In Figure 6, likewise indicated in Figure 2, the images comprise of a 176 x 234 pixels matrix with pixel size of 1mm x 1mm and they are contrast-brightness enhanced.

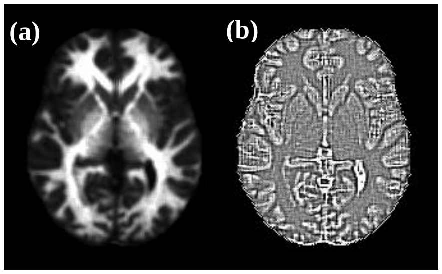


Figure 7. The image in (a) shows the Classic-Curvature and the image in (b) shows the Intensity-Curvature Functional. The original MRI is provided by the courtesy of the OASIS database [15].

Figure 7 shows the Classic-Curvature in (a) and the Intensity-Curvature Functional in (b). The images were obtained when fitting the bivariate cubic Lagrange polynomial and re-sampling with a misplacement (x, y) ≡ (0.01mm, 0.01mm) in (a) and a misplacement (x, y) ≡ (0.95mm, 0.95mm) in (b). The behavior of the two mathematical engineering tools is similar to the one showed in Figure 6. The Classic-Curvature of Figure 7a shows remarkable reproduction of the original MRI image features overall all of the anatomical structures. The Intensity-Curvature Functional seen in Figure 7b performs feature extraction, showing details of the MRI, which are not visible otherwise. And specifically, in both of (a) and (b) is highlighted the distinction between gray and white matter of the human brain. Figure 7b shows similarities with Figure 4b (also an Intensity-Curvature Functional image), with the exception that the third dimension seen as perpendicular to the image plane is not visible in Figure 7b. However, the level of details is more pronounced in Figure 7b than it is in Figure 4b, notwithstanding the contrast enhancement of the two images. Likewise the images in Figure 4, the images in Figure 7 have a pixels matrix size of 176 x 208 with pixel size of 1mm x 1mm. In Figure 7, the images in (a) and (b) are contrast-brightness enhanced.

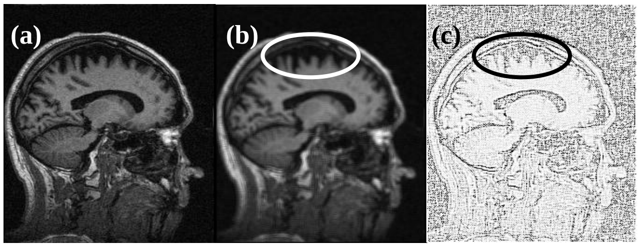


Figure 8. The image in (a) shows the original MRI, and (b) and (c) show the Classic-Curvature and the Intensity-Curvature Functional respectively. The effect of the Intensity-Curvature Functional is not as accentuated as the one seen in Figure 5c, and it is similar to the effect seen in Figure 5b.

Figure 8 shows the Classic-Curvature (see (b)) and the Intensity-Curvature Functional (see (c)) of the pathological MRI shown in (a). The subject was classified positive to the Clinical Dementia Rating (CDR) [10][12]. The Classic-Curvature image reproduces the original MRI shown in (a) with high level of details for what pertains to all of the anatomical structures and therefore highlights the shrinkage of cortical surface that can be seen in the regions inside the white ellipse in (b). As far as regards to the shrinkage of the cortical surface, the Intensity-Curvature Functional image seen in (c) shows that the intensity level is comparable to the noise level (see regions inside the black ellipse in (c)), thus adding confirmation to the observation made through the Classic-Curvature image. Both of the Classic-Curvature and the Intensity-Curvature Functional were obtained when fitting the bivariate cubic Lagrange polynomial and re-sampling of a misplacement (x, y) ≡ (0.01mm, 0.01mm). The pixels matrix size is 256 x 256 with pixel size 1mm x 1mm. The images were cropped to highlight the regions of interest and were contrast-brightness enhanced.

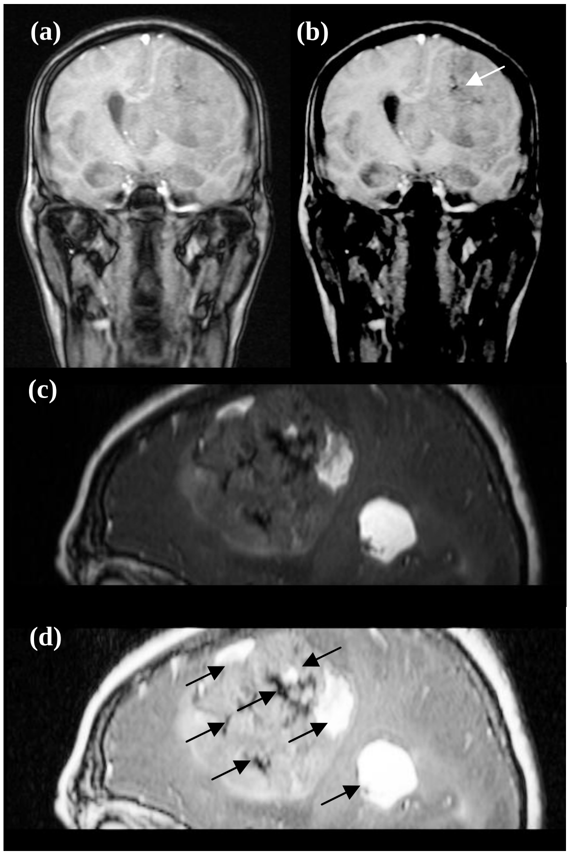


Figure 9. The images in (a) (contrast-brightness enhanced) and in (c) show the original MRI. The images in (b) and in (d) show the Intensity-Curvature Measure obtained fitting the data with the one-dimensional Sinc interpolation function.

Figure 9 shows in (a) and in (c) the original MRI with the tumor. Also, Figure 9 shows in (b) and in (d) the Intensity-Curvature Measure [1] obtained with the one-dimensional Sinc function. The interesting feature of the Intensity-Curvature Measure (see (b)) is the capability to highlight the tumor when extracting information from the original MRI seen in (a). Changing the brightness-contrast enhancement of Figure 9b yields an image, which is clearer than the one shown in (a), and which is the highlight of the tumor in its full spatial extent. The comparison between Figure 9a and 9b makes a clear and direct point, which is that the behavior of the mathematical engineering images is capable to add additional information to the original MRI (see Figure 9b: clearer contour line of the tumor and the dark spot indicated by the white arrow, which is presumably blood). In Figure 9b (contrast-brightness enhanced), the Intensity-Curvature Measure shows the capability to reveal the tumor with a perspective, which is different from the original image seen in Figure 9a. The tumor region is more clearly demarcated in (b) than it is in (a), whereas Figure 9d (contrast-brightness enhanced) demonstrates the capability of the

Intensity-Curvature Measure to focus on the fluids of the tumor as it is indicated through the arrows. The matrix size in Figures 9a and 9b is 512x512 pixels with 0.55mm x 0.55mm pixel size. The matrix size in Figures 9c and 9d is 512x512 pixels with 0.39mm x 0.39mm pixel size. The images in Figures 9a, 9b, 9c and 9d were cropped so to focus on the regions of interest. Similar behavior is observable in the mathematical engineering images shown by this piece of research. Figure 9d shows the contour line of the tumor and the fluids such as water and blood (see arrows). Both of the images seen in Figures 9b and 9d where obtained when re-sampling of 0.1mm along the x axis alone. The lesson learned is that the three mathematical engineering tools are able to extract additional and/or complementary information from MRI images. Future work should address the biomedical value of the information extracted from the MRI images through the methodology presented in this paper.

III. Discussion

This Section discusses on the effect of the contrast-brightness enhancement and also offers insights about the contribution of the works herein presented to the biomedical imaging processing literature stressing on the use of the mathematical engineering tools used to extract complementary and/or additional information from Magnetic Resonance Images of the human brain.

A. The Effect of the Brightness-Contrast Enhancement

The mathematical engineering images resulting from the original MRI: (i) Classic-Curvature, (ii) Intensity-Curvature Functional and (iii) Intensity-Curvature Measure have been object of brightness-contrast enhancement, whereas the original MRI brain images were not object of brightness-contrast enhancement (except for Figure 9a).

It can be argued that using the aforementioned enhancement in both the original MRI and the mathematical engineering images yields the same (or similar) result and thus the mathematical engineering images are not capable to add additional information to the original MRI (such possibility is explored in Figure 10). However, the capability of the mathematical engineering images of adding additional and/or complementary is supported in: (i) Figures 3, 4, 7 and 9b, and (ii) the following facts.

The first fact is that the mathematical engineering images, as widely observed in both of the cases herein reported and the cases that were reported elsewhere [1][3][17], present the characteristic of having pixel intensity values, which is quite different from the original brain images. Even with the large difference in pixel intensity values it is possible to set the same level of brightness-contrast enhancement for both of the original MRI and the mathematical engineering images. However, when the level of brightness-contrast enhancement is set the same, in the vast majority of the cases, different demarcation and appearance of the overall anatomical structure of the brain images was observed (see Figure 9a versus Figure 9b). Also, the aforementioned pixel intensity value difference makes it necessary the brightness-contrast enhancement of the mathematical engineering images so to view the content.

The second fact is consequential to the first one and is that the brightness-contrast enhancement is necessary to highlight the content of the mathematical engineering images so to reveal the additional and/or complementary information to the MRI. Indeed, through the mathematical engineering images it is possible to see: (i) the depth of the brain sulci (see Figure 3), (ii) the anatomical structure (see Figure 6a), (iii) the difference between gray and white matter (see Figure 1b and Figure 3a), (iv) the presence of fluids such as blood and water (see Figure 9d: the pathological image) and also the third dimension (see Figure 1c).

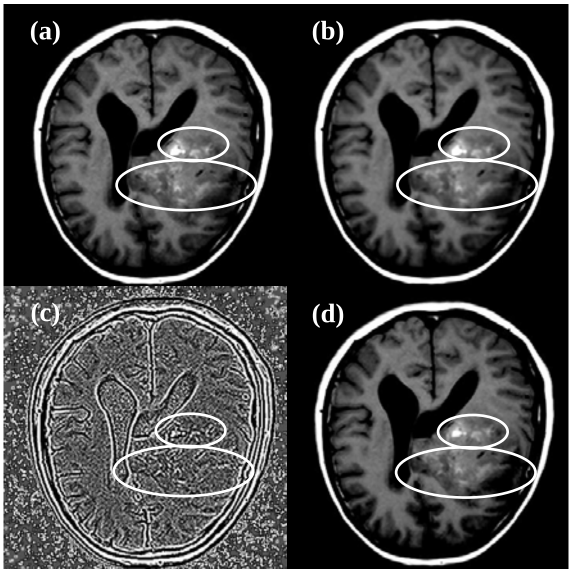


Figure 10. The image in (a) shows the original MRI with a tumor, whereas (b) shows the Classic-Curvature image, (c) shows the Intensity-Curvature Functional, and (d) shows the Intensity-Curvature Measure.

In order to investigate what happens when the contrast-brightness enhancement is set the same for both of the mathematical engineering images and the original MRI the following experiment was performed. Figure 10 reports the results of the experiment, which show that the Classic-Curvature (see Figure 10b) and the Intensity-Curvature Measure (see Figure 10d) images present almost the same characteristics of the original MRI, except for some blurring, which is visible because of the mathematical processing. Instead, the Intensity-Curvature Functional (see Figure 10c) shows details that are not observable in the original MRI (see inside the white ellipses). The images in Figure 10b and Figure 10c were obtained when fitting to the MRI data the bivariate cubic Lagrange model function when re-sampling of 0.1mm along both x and y directions, whereas the image in Figure 10d was obtained when fitting the one-dimensional Sinc model function when re-sampling of 0.1mm along the x direction. The matrix size of the images in Figure 10 is 512x512 with pixel size of 0.49 mm x 0.49 mm. The images in (a), (b) and (d), have been set to the same brightness-contrast adjustment. In (d) the Intensity-Curvature Functional shows a complementary perspective to the images seen in (a) and (b), highlighting the structure of the tumor fluids such as blood and water. The images were cropped to highlight the regions of interest and are all contrast-brightness enhanced.

The significance of Figure 10 is that the mathematical engineering images are capable to show the anatomical structure of the human brain likewise the original MRI does. This fact is positive to the research question of the herein presented work, which investigates whether the mathematical engineering images add complementary information to the MRI. Also, Figure 10c shows that the Intensity-Curvature Functional presents details which are not visible in the original MRI. Such fact is in favor to the aforementioned research question. Hence, the contrast-brightness enhancement, is not a confounding factor, it is indeed a requirement for the extraction of additional and/or complementary information from the MRI of the human brain because at the least the mathematical engineering images can reproduce the same level of details of the MRI of the human brain (see Figure 10).

B. *The Contribution to Biomedical Image Processing*

A well-defined novel formulation of three specific mathematical engineering instruments has been conceived [2]-[4]. The novel formulation provides the solution of the biomedical signal processing problem, which consists in extracting additional information from the Magnetic Resonance Imaging (MRI) signal of the human brain.

The three mathematical engineering instruments are called: (i) Classic-Curvature, (ii) Intensity-Curvature Functional and (iii) Intensity-Curvature Measure. The three math instruments make use of the second order derivatives of the model function fitted to the data. The aforementioned instruments makes it possible to re-image the Magnetic Resonance Imaging (MRI) image data of the human brain into three novel domains where there exists features that would not be otherwise observable in the original MRI images.

The research herein presented has significance in the field of biomedical signal processing and more generally in diagnostic radiology because brings to the attention of the reader the existence of three novel domains. The novel domains have been revealed through the use of conceptual forms descending from one main signal processing technique, which is that of the calculation of the Classic-Curvature [1]-[3]. In fact, the calculation of the Classic-Curvature enables also the calculation of the Intensity Curvature Functional and the Intensity-Curvature Measure. The work herein presented demonstrates the feasibility of the calculation of the Classic-Curvature, the Intensity-Curvature Functional and the Intensity-Curvature Measure from the two-dimensional MRI of the human brain both in normal and pathological cases. The calculation of the Intensity-Curvature Functional in three dimensions is also possible [17].

IV. CONCLUSION

This paper considered the problem of the three mathematical engineering tools that are used to provide useful information, which is not readily observable into the original MRI images. This research also provides results which evaluate the performance of the proposed mathematical engineering tools. The results show that the Classic-Curvature image reproduces the original MRI image with high level of details and both of the Intensity-Curvature Functional and the Intensity-Curvature Measure performs feature extraction showing details of the MRI which are not visible otherwise. The advantage provided through the re-sampling process is indeed a fact which gives an immense number of possible Classic-Curvature, Intensity-Curvature Functional and Intensity-Curvature Measure images. Also, it provides the freedom to choose images which give best result in diagnostic practice. The mathematical models rely on software code implementing complex math formulas. Due to the originality of the research here presented it is not possible to compare our results with previous research findings. However, since we provide the software free of charge, the mathematical engineering tools are easily available to the scientific community.

ACKNOWLEDGMENT

Hereto reporting findings that benefit from OASIS data [16], we are due to mention the following grant numbers: P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382 R01 MH56584 [16]. The MRI images shown in Figure 1a (© IGI Global Publisher) and Figure 2b (© IGI Global Publisher) are reprinted from [2]. The MRI image shown in Figure 2a (© Carlo Ciulla) is reprinted from [3].

REFERENCES

[1] C. Ciulla, “On the calculation of the signal-image classic-curvature: A second order derivatives based approach”, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol. 2, no. 4, 2013, pp. 158-165.

[2] C. Ciulla, “Improved signal and image interpolation in biomedical applications: The case of magnetic resonance imaging (MRI) ”, Medical Information Science Reference - IGI Global Publisher, Hershey, PA, U.S.A., 2009.

[3] C. Ciulla, “Signal resilient to interpolation: An exploration on the approximation properties of the mathematical functions”, CreateSpace Publisher, U.S.A., 2012.

[4] C. Ciulla, “On the signal-image intensity-curvature content”, International Journal of Image, Graphics and Signal Processing, vol. 5, no. 5, 2013, pp. 15-21.

[5] Y. Cha and S. Kim, “The error-amended sharp edge (EASE) scheme for image zooming’, IEEE Transactions on Image Processing”, vol. 16, no. 6, 2007, pp. 1496-1505.

[6] J. Prewitt, “Object enhancement and extraction”, In: Picture Processing and Psychopictorics, B. Lipkin and A, Rosenfeld, Eds. New York: Academic, 1970, pp. 75-149.

[7] S.K. Lele, “Compact difference schemes with spectral-like resolution”, Journal of Computational Physics, vol. 103, 1992, pp. 16-42.

[8] H. Farid and E.P. Simoncelli, “Differentiation of discrete multidimensional signals”, IEEE Transactions on Image Processing, vol. 13, no. 4, 2004, pp. 496-508.

[9] R.L. Buckner et al., “A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: Reliability and validation against manual measurement of total intracranial volume”, Neuroimage, vol. 23, no. 2, 2004, pp. 724-738.

[10] A. F. Fotenos, A. Z. Snyder, L. E. Girton, J. C. Morris, and R. L. Buckner, “Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD”, Neurology, vol. 64, 2005, pp. 1032-1039.

[11] D. S. Marcus et al., “Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults”, Journal of Cognitive Neuroscience, vol. 19, no. 9, 2007, pp. 1498-1507.

[12] J. C. Morris, “The clinical dementia rating (CDR): current version and scoring rules”, Neurology, vol. 43, no. 11, 1993, pp. 2412b-2414b.

[13] E. H. Rubin et al., “A prospective study of cognitive function and onset of dementia in cognitively healthy elders”, Archives of Neurology, vol. 55, no. 3, 1998, 2001, pp. 395-401.

[14] Y., Zhang, M. Brady, and S. Smith, “Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm”, IEEE Transactions on Medical Imaging, vol. 20, no. 1, 2001, pp. 45-57.

[15] OASIS database <http://www.oasis-brains.org/> 2014.05.15

[16] C. Ciulla and F. P. Deek “The Sub-Pixel Efficacy Region of the Bivariate Linear Interpolation Function”, International Journal of Computer Applications in Technology, vol. 49, nos. 3/4, May, 2014, pp.270–281.

[17] C. Ciulla, “The intensity-curvature functional of the trivariate cubic Lagrange interpolation formula”, International Journal of Image, Graphics and Signal Processing, vol. 5, no. 10, 2013, pp. 36-44.

A New Blind Equalization Algorithm for M-PSK Constellations

Ali Ekşim and Serhat Gül
Center of Research for Advanced Technologies of Informatics and Information Security (TUBITAK-BILGEM)
41470, Gebze, Kocaeli, Turkey
{ali.eksim, gul.serhat}@tubitak.gov.tr

***Abstract*—In this paper, we propose an equalization algorithm for *M-PSK* constellations that greatly improves the convergence features and reduces steady-state error rate of the conventional Constant Modulus Algorithm (CMA). The proposed algorithm introduces a buffer and multiple step-size decision layers to the existing Variable Step Size Modified Constant Modulus Algorithm (VSS-MCMA) equalizer. The buffer is employed to combat the convergence issue while the additional layers have been introduced to improve sensitivity of the step size in both the convergence state and the steady-state. Computer simulations reveal that the proposed algorithm has better convergence rate than the VSS-MCMA and the CMA.**

Keywords—Blind equalization; constant modulus algorithm; step size

I. Introduction

In wireless communications, one of the most important transmission problems is the channel distortion. Channel distortion leads to InterSymbol Interference (ISI) between transmitted symbols. There have been many blind equalization techniques to combat this effect. Constant modulus algorithm, originated by Godard [1] and Treichler and Agee [2], is the most popular equalization technique among all blind equalization methods. As all blind equalizers, constant modulus algorithm works in absence of the training sequence. In the algorithm, step size is a crucial parameter to determine the convergence speed and steady-state error rate. A small chosen step size will result in a low steady-state error rate. However, convergence will be slow. Conversely, a large value of step size will result in a faster convergence yet a higher steady-state error rate. Therefore, the Constant Modulus Algorithm (CMA) has a trade-off between these two criteria. Another drawback of the algorithm is that it is unable to correct phase rotations induced by the channel [3].

Oh and Chin propose the Modified Constant Modulus Algorithm (MCMA), which resolves phase rotation problem of the CMA. This is achieved by minimizing two cost functions, which are separately computed for real and imaginary parts of input signal. The algorithm applies both equalization and phase correction.

Variable Step Size Modified Constant Modulus Algorithm (VSS – MCMA) [4] is an algorithm that applies phase correction and makes the step size more sensitive. In this algorithm, a circular area is defined around each likely-transmitted symbol in the constellation and the step size is changed between two values according to whether equalizer output symbol is located inside an area or not.

Many other algorithms [5]-[17] have been proposed to combat the problems encountered in CMA. Zarzoso and Comon [5] suggested an algorithm that finds optimal step size in each update operation. Tugcu et al. [6] employ cross correlation between channel output and the error signal to overcome the convergence problem. Song et al. [7] employ a signal steering vector and its oblique projection for updating filter coefficients in order to avoid signal steering vector mismatches. Lin and Lee [8] introduce an algorithm that finds a gain factor by the least-mean-squares method when updating filter coefficients to avoid gradient noise amplification problem. Demir and Ozen [9] proposed a new algorithm in which autocorrelation of error signal is used for updating filter coefficients. Gao and Qiu [10] employ a momentum term and autocorrelation of error signal for updating filter coefficients. The additional momentum term improves the convergence rate. Li et al. [11] utilize singular value decomposition of input signal to obtain a new step size in order to improve the convergence rate. A new update equation is proposed by Abrar and Nandi [12] to improve the convergence rate. A nonlinear estimate of error signal is used for updating the equalizer coefficients, and a novel deterministic optimization criterion is given. Ikhlef et al. [13] employ the prewhitening technique and the complex Givens rotations to improve signal to noise and interference ratio performance. Yan et al. [14] employ nonlinear transformation of error signal to suppress the alpha-stable noise. Nassar and Nahal [15] recently proposed Exponentially weighted step-size recursive Least Squares Constant Modulus Algorithm (EXP-RLS-CMA), which can be considered as the combination of the conventional CMA and the exponentially weighted step-size recursive least squares algorithm. The EXP-RLS-CMA provides higher convergence rate than the conventional CMA at minimum mean-squared-error. Liyi et al. [16] introduce a new variable step size algorithm in which a nonlinear function of error signal is employed to calculate the step size in each symbol period. Baofeng et al. [17] employ cross-correlation between the input signal and the error signal to control the step size for a better convergence rate.

In this work, we have generalized the VSS-MCMA to multi-layered case in order to make the step size more sensitive. Moreover, we have added a buffer to the equalizer to overcome the convergence problem.

This paper is organized as follows. We have introduced the general transmission model and system models of the CMA, MCMA and VSS-MCMA equalizers in Section II. In Section III, we have presented our Buffered Multi-layered Modified Constant Modulus Algorithm (BML-MCMA). We have examined the simulation results of the proposed algorithm and the other two algorithms in Section IV. Finally, we have drawn the conclusions in Section V.

II. SYSTEM MODELS

Consider a baseband transmission model where the received signal can be written as

$$r(n)=\sum_{k=0}^{L-1}h(k)s(n-k)+u(n) \qquad (1)$$

where $h(n)$ is the channel's impulse response of length L, $s(n)$ is the transmitted complex baseband symbol at time n and $u(n)$ is additive white noise. If the received signal is fed to the equalizer, the output is

$$z(n)=R^{H}W(n)\,. \qquad (2)$$

In the above equation, $W(n)$ is the equalizer tap vector of length-N, which is defined as $W(n)$ = $[w_0(n), w_1(n), \dots , w_{N-1}(n)]^T$. $R(n)$ is the tapped delay line vector of received signal $r(n)$ and it is defined as $R(n)$ = $[r(n), r(n\text{-}1), \dots , r(n\text{-}N\text{+}1)]^T$.

A. Constant Modulus Algorithm

The CMA developed by Godard [1] and Treichler [2] is a stochastic gradient-based algorithm. The cost function is given by

$$J(n)=E\left[\left(\gamma-|z(n)|^2\right)^2\right] \qquad (3)$$

where parameter γ is given by $\gamma \triangleq E\{|s(n)|^4/|s(n)|^2\}$. Here, $E\{.\}$ denotes expectation. In this algorithm, equalizer coefficients are updated by minimizing the cost function in (3). The update rule is

$$W(n+1)\ =\ W(n)+\mu\vec{g} \qquad (4)$$

where μ is the step size parameter, $\vec{g}$ is the gradient vector defined as $\vec{g}=\nabla J(n)=e(n)R^{*}(n)$. The error signal $e(n)$ is given by $e(n)=z(n)\left(\gamma-|z(n)|^2\right)^2$. Expanding right-hand side of the expression in (4) using the above definitions yields

$$W(n+1)=W(n)+\mu z(n)\left(\gamma-|z(n)|^2\right)^2 R^{*}(n)\,. \qquad (5)$$

B. Modified Constant Modulus Algorithm

MCMA blind equalization algorithm in [3] applies modifications to CMA in calculation of cost function. Computations are performed using real and the imaginary parts of equalizer output separately. The purpose of separating the real and imaginary parts is to correct phase rotations encountered in CMA. The cost function for MCMA has the form

$$J(n)=J_R(n)+J_I(n)\,. \qquad (6)$$

In (6), $J_R(n)$ and $J_I(n)$ are cost functions of real and imaginary parts of equalizer output, respectively, and they are defined as

$$J_R(n)=\left(|z_R(n)|^2-\gamma_R\right)^2 \qquad (7)$$

$$J_I(n)=\left(|z_I(n)|^2-\gamma_I\right)^2. \qquad (8)$$

Here, γ_R and γ_I are two parameters for real and imaginary parts of transmitted symbol respectively and they are defined as $\gamma_R = E\{|s_R(n)|^4/|s_R(n)|^2\}$ and $\gamma_R = E\{|s_R(n)|^4/|s_R(n)|^2\}$ where $s(n) = s_R(n) + js_I(n)$.

C. Variable Step Size Modified Constant Modulus Algorithm

The VSS – MCMA algorithm proposed in [4] is an improvement on MCMA. The algorithm employs step-size adaptation to boost the performance in both the convergence state and the steady state. The change of step size depends on the regions defined in the signal space.

Let the signal space contain M regions, namely, a circular area with radius l is placed around each of total M likely transmitted symbols. One of two step size parameters is selected according to whether the equalizer output is located inside the circular area around the nearest symbol or not. In other words

$$\mu=\begin{cases}\mu_{VSS,0}, & z(n)\notin B_i\ ,\ \ i=1,2,\dots,M\\ \mu_{VSS,1}, & z(n)\in B_i\ ,\ \ i=1,2,\dots,M\end{cases} \qquad (9)$$
$$\mu_{VSS,0}>\mu_{VSS,1}.$$

The above equation indicates that if equalizer output is not located inside the area B_i, the algorithm selects the larger step size, $\mu_{VSS,0}$. If not, the algorithm selects the smaller one, $\mu_{VSS,1}$.

III. BUFFERED MULTI-LAYERED MODIFIED CONSTANT MODULUS ALGORITHM

In the proposed system, existing VSS – MCMA algorithm is generalized to a multi-layered system and parallel buffer-delay elements are added before the equalizer. The purpose of the changes is to overcome the convergence problem and lower the steady state error rate. Block diagram of the proposed system is shown in Fig. 1. In the system, initially, first *S* baud segment of input signal *r*(*n*) is used for updating the equalizer coefficients only. At the end of *S* baud periods, namely, at time ST_S, thanks to the control element, the system turns off the counter and switches to delay output. This time, input signal is fed into the equalizer through the delay element. Although a delay of *S* baud periods is introduced to the signal, because the equalizer coefficients are updated previously, the convergence problem is removed on a large scale.

To begin analysis of the system, let the signal space contain *D* regions around every symbol, namely, around every symbol, *D* circular layers are constructed. This is depicted in Fig. 2, where *D* has the value of 2. Here, $l_{i,j}$ (i = 1,2, … , D, j = 1,2,…, M) denotes the radius of *i*th layer around *j*th symbol. Let us denote radius of the outmost layer by *l* and the distance between two nearest symbols by *p*. Then, the restriction to the size of outmost layer is l<0.5p. The step size is selected according to whether the equalizer output is located in a layer around nearest symbol or not. The general selection rule can be expressed as

$$\mu = \begin{cases} \mu_0 \quad , & l_{D,j} < \left\| z(n) - s_j \right\| < \dfrac{p}{2} \\ \mu_1 \quad , & l_{(D-1),j} < \left\| z(n) - s_j \right\| < l_{D,j} \\ \cdot & \\ \cdot & \\ \cdot & \\ \mu_{D-1} \quad , & 0 < \left\| z(n) - s_j \right\| < l_{1,j}. \end{cases} \tag{10}$$

In (10), s_j is the *j*th transmitted symbol in the constellation and *z*(*n*) is the output signal of the equalizer at time instant *n*. Relationship between the step size parameters is $\mu_0 > \mu_1 > \ldots > \mu_{D-1}$.

IV. NUMERICAL RESULTS

In this section, we explore the convergence rate performance of the proposed algorithm through computer simulations. We also compare performance of the proposed algorithm with the conventional CMA and the VSS – MCMA.

In the simulations, we have used the 2-tap complex channel $h(n)$ = [1 + 0.5$\delta(n$-1)] + j[1 + 0.4$\delta(n$-1)] [13] and the 3-tap Proakis B-channel $h(n)$ = 0.407 + 0.815$\delta(n$-1) + 0.407$\delta(n$-2) [18], where $\delta(n)$ is the Kronecker delta function. The channel is normalized such that the total power is unit watts. Equalizer tap number is selected as 11. For the 2-tap complex channel, all of the taps are initialized to zero except the first tap is initialized to one. For Proakis-B channel, all of the taps are initialized to zero except the center tap is initialized to one. Additive noise is assumed to be zero-mean complex white Gaussian noise. Transmitted signal length *N* is chosen as 10^6 symbols, which is assumed to be long enough to examine the performance of the three systems. Number of buffer samples *S* is chosen as 10^4 for the 2-tap complex channel case and 5×10^4 for the Proakis B-channel case.

We present the performance of the three equalizers through constellations of equalizer outputs. Constellations of outputs of the three equalizers at QPSK, 8-PSK, 16-PSK and 32-PSK under the 2-tap complex channel, are given in Fig. 3, Fig. 4, Fig. 5 and Figs. 6 and 7, respectively. SNR values at which the simulations are performed, are chosen as 15 dB, 20 dB, 30 dB and 40 dB respectively. Optimum values of μ and *l* parameters of the three equalizers are obtained heuristically.

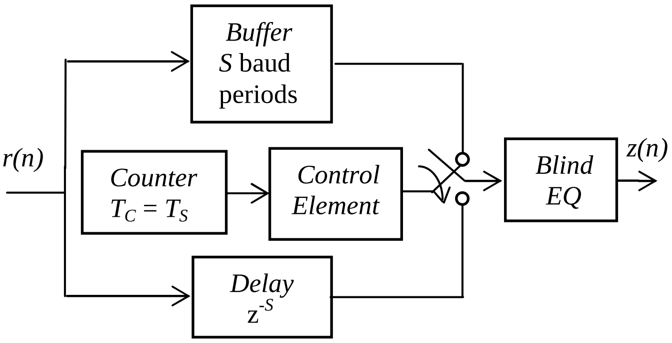


Fig. 1. Block diagram of the proposed BML – MCMA system

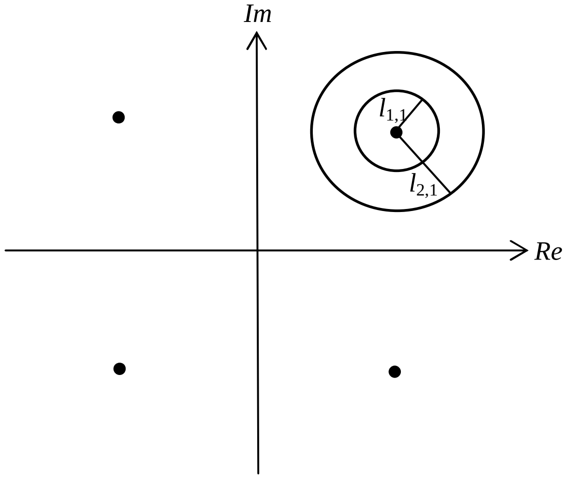


Fig.2. Step-size decision layers (*D* = 2, QPSK).

In Fig. 3, it can be seen that BML-MCMA has the lowest Mean-squared Error (MSE) by having less scattered symbols than the two other. In the case of VSS-MCMA, a clear QPSK constellation is not visible. Therefore, the VSS-MCMA has the highest MSE and Symbol Error Rate (SER).

From Fig. 4, one can note that the BML-MCMA outperforms the two other. The employment of buffer and multiple step size layers results in a better output constellation. Because there is no buffer used in conventional CMA and VSS-MCMA, number of highly scattered symbols are much greater than those of the BML-MCMA. The conventional CMA has a better constellation than VSS-MCMA.

Fig. 5 clearly indicates that BML-MCMA has the best performance by having the clearest constellation. Here, again, there are large number of highly scattered symbols in constellations of CMA and VSS-MCMA. The BML-MCMA overcomes this problem thanks to its buffer element and its output constellation yields better overall MSE than those of the other two.

Fig. 6 and Fig. 7 show that the BML-MCMA has the best MSE performance and convergence rate. Conventional CMA and the VSS-MCMA have many highly-scattered symbols and they cannot yield a clear constellation. Consequently, these two equalizers have much lower convergence rate.

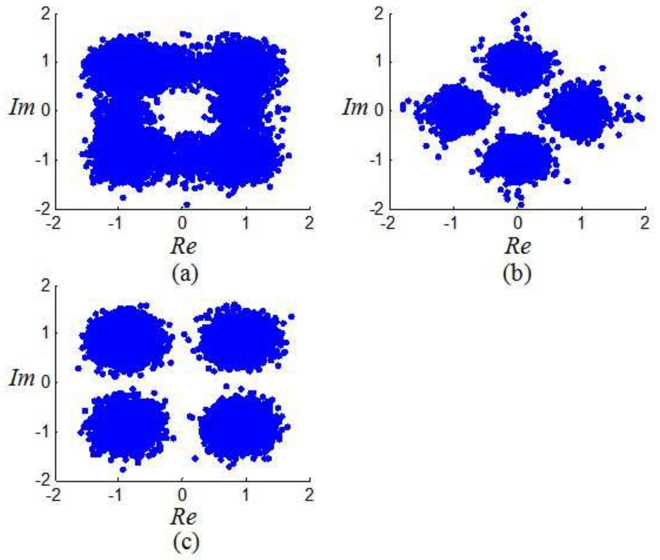


Fig. 3. Constellation diagrams of first 20000 elements of equalizer output at QPSK modulation. (a) VSS-MCMA, (b) CMA, (c) BML-MCMA

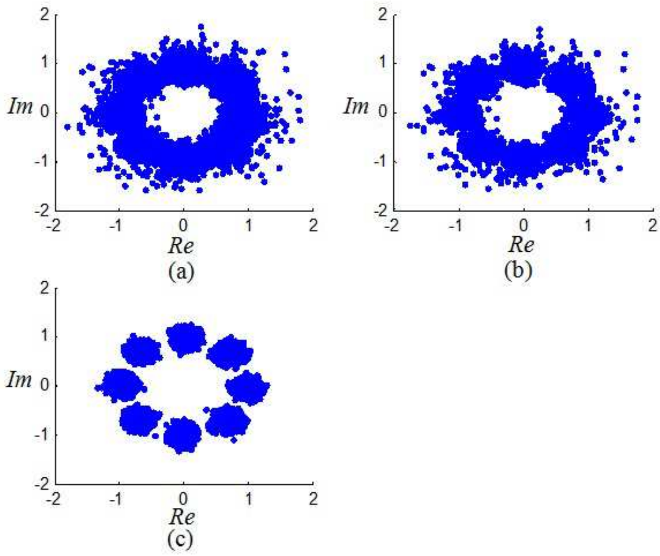


Fig. 4. Constellation diagrams of first 20000 elements of equalizer output at 8-PSK modulation. (a) VSS-MCMA, (b) CMA, (c) BML-MCMA

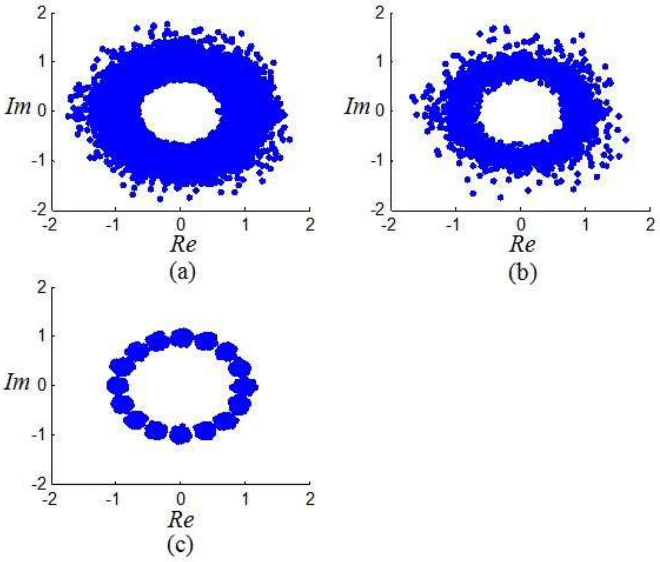


Fig. 5. Constellation diagrams of first 20000 elements of equalizer output at 16-PSK modulation. (a) VSS-MCMA, (b) CMA, (c) BML-MCMA

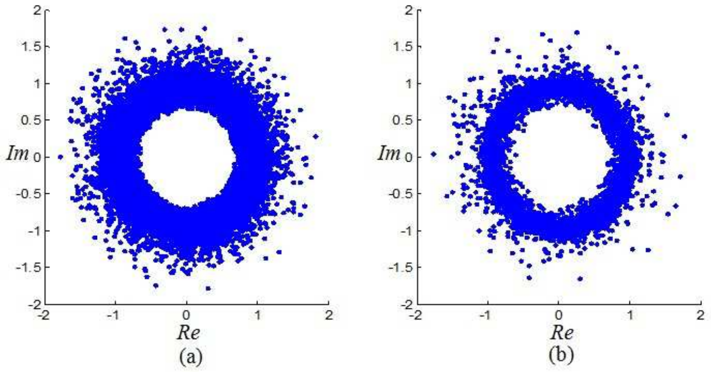


Fig. 6. Constellation diagrams of first 20000 elements of equalizer output at 32-PSK modulation. (a) VSS-MCMA, (b) CMA

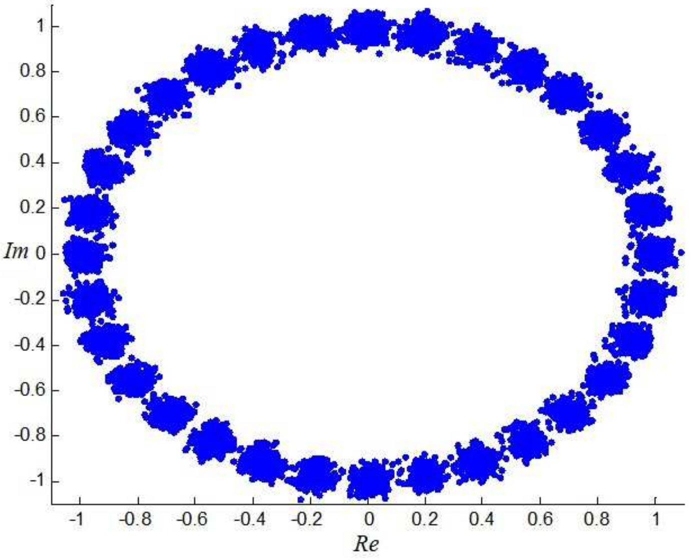


Fig. 7. Constellation diagram of first 20000 elements of BML-MCMA equalizer output at 32-PSK modulation.

Fig. 8 shows the constellation of first 20000 output symbols of equalizer at 8-PSK modulation under Proakis B-channel. Since the Proakis B-channel is a harsh channel, we have increased the buffer size to 50000 symbols. From Fig. 8, it can be observed that BML-MCMA outperforms the two other in convergence rate. Conventional CMA takes the second place and VSS-MCMA again has the lowest convergence rate.

In the second part of the simulations, we have investigated the effect of buffer size *S* on the convergence rate of the BML-MCMA equalizer through the constellations obtained for various buffer size values. Fig. 9 shows the constellations of first 20000 elements of equalizer outputs for *S* = 10000, 25000 and 50000 symbols, respectively. The modulation is 8-PSK and we have used the Proakis B-channel. The SNR value is 16 dB. Fig. 9 shows that using only the first 50000 symbols of the input, the BML-MCMA equalizer reduces the convergence problem greatly and results in a clear constellation. For a buffer size of 25000 symbols, the equalizer significantly reduces the convergence problem so it can resolve the 8-PSK symbols. Due to the harsh Proakis B-channel, a buffer size of 10000 symbols is not enough to eliminate the convergence problem.

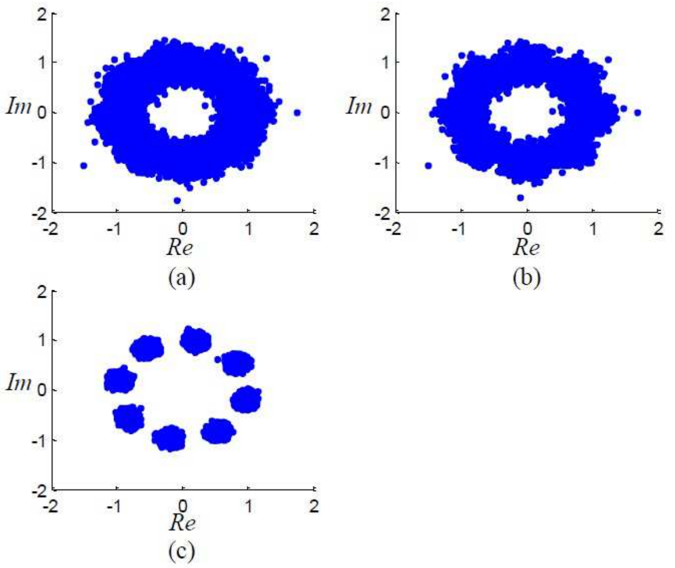


Fig. 8. Constellations of first 20000 elements of equalizer outputs at 8-PSK modulation. (a) VSS-MCMA, (b) CMA, (c) BML-MCMA

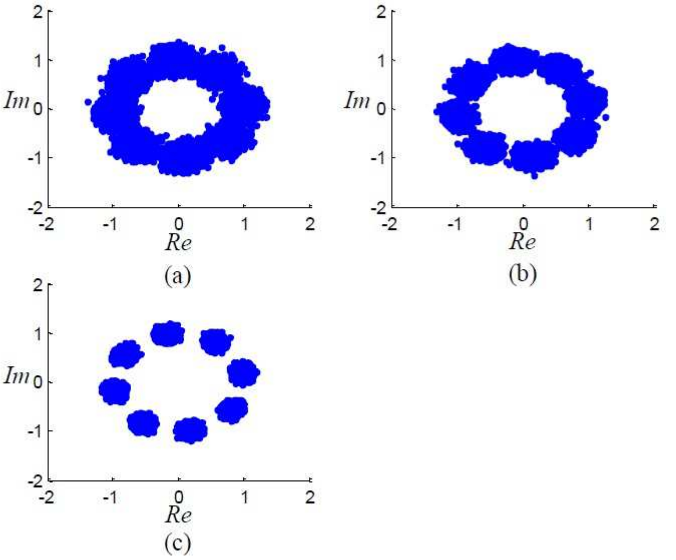


Fig. 9. Constellations of first 20000 elements of equalizer output under different buffer size values. (a) $S = 10000$, (b) $S = 25000$, (c) $S = 50000$ symbols

V. CONCLUSION

Despite there is no need for a training sequence, conventional CMA and VSS-MCMA have convergence problems. In this paper, we have proposed our BML-MCMA algorithm to eliminate the convergence problems and further improve MSE performance of VSS-MCMA. We have presented numerical results which indicate that through using a buffer, the proposed algorithm resolves convergence problem greatly. Furthermore, the proposed algorithm has better MSE performance. Although using a buffer introduces some initial delay to the equalizer, considering the overall performance, BML-MCMA is a good alternative to the existing blind equalization algorithms.

VI. REFERENCES

[1] D. N. Godard, “Self-recovering equalization and carrier tracking in two-dimensional data communication systems,” IEEE Trans. on Commun., vol. 28, no. 11, Nov. 1980, pp. 1867-1875.

[2] J. R. Treichler and B. G. Agee, “A new approach to multipath correction of constant modulus signals,” IEEE Trans. Acoust., Speech, Signal Processing, vol. 31, no. 2, Apr. 1983, pp. 459–472.

[3] K. N. Oh and Y. O. Chin, “Modified constant modulus algorithm: Blind equalization and carrier phase recovery algorithm,” IEEE International Conference on Communications, vol. 1, June 1995, pp. 498 -502.

[4] Wei. X, X. Yang, and Z. Zhang, “A variable step size algorithm for blind equalization of QAM signals,” Progress In Electromagnetic Research Symposium Proceedings, July 2010, pp. 271-275.

[5] V. Zarzoso and P. Comon, “Optimal step-size constant modulus algorithm,” IEEE Transactions on Communications, vol. 56, no. 1, Jan. 2008, pp. 10-13.

[6] E. Tugcu, F. Cakir, and A. Ozen, “A novel variable step size constant modulus algorithm employing cross correlation between channel output and error signal,” 35th International Conference on Telecommunications and Signal Processing (TSP), July 2012, pp. 678-683.

[7] X. Song, J. Wang, Q. Li, and H. Wang, “Robust constrained constant modulus algorithm for signal steering vector mismatches,” 8th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM), vol. 1, Sept. 2012, pp. 1-4.

[8] J. C. Lin and L. S. Lee, “A modified blind equalization technique based on a constant modulus algorithm,” International Conference on Communications, IEEE Conference Record, vol. 1, Jun 1998, pp. 344-348.

[9] M. A. Demir and A. Ozen, “A novel variable step size constant modulus algorithm based on autocorrelation of error signal for blind equalization,” 34th International Conference on Telecommunications and Signal Processing (TSP), Aug. 2011, pp. 500-504.

[10] Y. Gao and X. Qiu, “A new variable step size CMA blind equalization algorithm,” 24th Chinese Control and Decision Conference (CCDC), 23-25 May 2012, pp. 315-317.

[11] G. Li, L. Ning; G. Yan, and Z. Jiongpan, “Convergence behavior of the constant modulus algorithm controlled by special stepsize,” 6th International Conference on Signal Processing, vol. 1, Aug. 2002, pp. 390-392.

[12] S. Abrar and A.K. Nandi, “An adaptive constant modulus blind equalization algorithm and its stochastic stability analysis,” IEEE Signal Processing Letters, vol. 17, no. 1, Jan. 2010, pp. 55-58.

[13] A. Ikhlef, K. Abed-Meraim, and D. Le Guennec, “On the constant modulus criterion: a new algorithm,” IEEE International Conference on Communications (ICC), May 2010, pp. 1-5.

[14] K. Yan, H. C. Wu, D. Xu, and S. S. Iyengar, “Novel robust blind equalizer for QAM signal using iterative weighted-least-mean-square algorithm,” IEEE Global Telecommunications Conference (GLOBECOM), 2010, pp. 1-6.

[15] A. M. Nassar and W. E. Nahal, “New blind equalization technique for constant modulus algorithm (CMA),” IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR), 8-10 June 2010, pp. 1-6.

[16] Z. Liyi, C. Lei, and S. Yunshan, “Variable Step-size CMA blind equalization based on non-linear function of error signal,” WRI International Conference on Communications and Mobile Computing (CMC), vol. 1, 6-8 Jan. 2009, pp. 396-399.

[17] Z. Baofeng, Z. Jumin, and L. Dengao, “A new variable step-size constant modulus blind equalization algorithm”, International Conference on Artificial Intelligence and Computational Intelligence (AICI), vol. 3, 23-24 Oct. 2010, pp. 289-291.

[18] J. G. Proakis, Digital Communications, 3rd. ed., McGraw-Hill, 1995.

A Single-Threshold Model for Handoff Traffic Analysis in Cellular CDMA Networks

Vassilios G. Vassilakis
Dept. of Electronic Engineering
University of Surrey
Guildford, U.K.
e-mail: v.vasilakis@surrey.ac.uk

Ioannis D. Moscholios
Dept. of Informatics and Telecommunications
University of Peloponnese
Tripolis, Greece
e-mail: idm@uop.gr

Michael D. Logothetis
WCL, Dept. of Electrical and Computer Engineering
University of Patras
Patras, Greece
e-mail: mlogo@upatras.gr

Michael N. Koukias
WCL, Dept. of Electrical and Computer Engineering
University of Patras
Patras, Greece
e-mail: koukias@upatras.gr

***Abstract*—Small cells are expected to play an important role in future mobile networks. In such environments, proper handling of handoff traffic is of major importance. In this paper, we propose the single-threshold model for the analysis of handoff traffic in cellular CDMA networks. Based on this model, we are able to determine analytically the uplink blocking probabilities of handoff and new calls. This is done by describing the CDMA system as a Discrete-Time Markov Chain and by deriving an efficient recursive formula for the calculation of system state probabilities. The proposed analytical model is verified via simulation studies.**

Keywords—handoff; cdma; call blocking probability; recursive formula.

I. Introduction

The Code Division Multiple Access (CDMA) techniques have been used in the current generation mobile networks and are expected to play an important role in future 5G networks. Some of the advantages of CDMA-based techniques over other competing technologies include enhanced security, efficient frequency spectrum utilization, and improved signal quality.

According to the traditional cellular model, the geographical area is divided into cells, and each of them is controlled by a Base Station (BS). Different BSs communicate with each other through the core network (usually fixed and wired). Although, in future mobile networks it is envisioned the introduction of intelligent BSs that will be able to communicate directly with each other for both signalling and data traffic [1]. Mobile Users (MUs), located in the same or different cells, communicate with each other through the corresponding BSs. The communication link from MUs to BS is referred to as uplink, whereas the communication link from BS to MUs is referred to as downlink. Due to non-orthogonality of the CDMA codes, a new MU arriving to a cell will cause interference to other MUs in the same and neighbouring cells. Therefore, Call Admission Control (CAC) is performed upon a call arrival, in order to protect the Quality-of-Service (QoS) of existing MUs. This may result in the blocking of the newly arriving call.

In this work, we consider two types of call blocking: new-call blocking and handoff-call blocking. The first type refers to the call blocking upon the initial connection establishment, whereas the second type refers to the blocking of already accepted in-service calls when they move from one cell to another. The procedure of moving between neighbouring cells, while a call is in progress, is called handoff. The CAC policy is expected to guarantee that the handoff-call blocking probability will be significantly lower than that of the new-call blocking.

In this paper, we model a cellular CDMA system as a Discrete-Time Markov Chain (DTMC). Our analysis is based on the classical Erlang Multirate Loss Model (EMLM) [2], [3] and its extension for single-threshold model [4], [5], which has been proposed for wired connection-oriented networks. We extend [2]-[5] by considering the handoff traffic and soft network capacity of CDMA systems.

This paper is organized as follows. In Section II, we present the literature review. In Section III, we describe our proposed model for cellular CDMA systems. In Section IV, we present a detailed calculation of local blocking probabilities. In Section V, we derive equations for an efficient calculation of call blocking probabilities. In Section VI, we study the performance of the proposed approach by means of computer simulations. We conclude and discuss our future work in Section VII.

II. Literature Review

Many important teletraffic models have been proposed for the determination of *new-call* blocking probabilities in cellular CDMA networks [6]-[18]. In [6], the call blocking calculation in the uplink of a W-CDMA cell is based on an extension of the EMLM. The authors assume that calls arrive in the system according to a Poisson process. This work was extended in [7], by incorporating elastic and adaptive traffic, and in [8], [9] by considering a quasi-random call arrival process. In [10], an efficient CAC scheme for CDMA systems has been proposed and evaluated. In [11], the authors propose an analytical model for multi-service cellular networks servicing multicast connections. An extension of [8] has been proposed in [12] to model elastic and adaptive traffic. A different approach that includes interference cancellation has been proposed in [13]-[15]. In [16], the authors evaluate the performance of W-CDMA systems with different QoS requirements. In [17], a teletraffic model for a W-CDMA cell with finite number of channels and finite number of traffic sources is presented. This model has been extended in [18] to provide equalization of call congestion probabilities among different service-classes.

Some of the aforementioned models have been extended for the analysis of *handoff-call* blocking probabilities [19]-[22]. In [19], a model for W-CDMA systems with a soft handoff mechanism has been proposed. In [20], the model of [8] has been extended for the calculation of handoff-call blocking probabilities. In [21] and [22], the model of [6] has been enhanced with a CAC for handoff traffic. While most of the works concentrate on the uplink, a few papers study the downlink of CDMA systems as well [23], [24].

In this paper, we concentrate on the uplink of cellular CDMA systems and handoff traffic. In particular, we extend [21] by enabling two contingency bandwidth requirements of the arriving calls. If the system is heavily loaded (above a predefined threshold), then the call will request less bandwidth compared with the case of lightly loaded systems.

III. MODEL DESCRIPTION

Consider a CDMA system that supports K independent service-classes. We examine a reference cell surrounded by neighbouring cells in the uplink direction (calls from MUs to BS).

The following QoS parameters characterize a service-class k (k=1,…, K) new call:

- $R_{k,N}$: Transmission bit rate.
- $(E_b / N_0)_{k,N}$: Bit error rate (BER) parameter.

The offered traffic-load (in erl) of service-class k new calls is Poisson and denoted as $a_{k,N}$. For our analysis, we express later in the paper the different service's QoS requirements as different resource/bandwidth requirements.

In a similar way, the QoS parameters of a service-class k handoff call are defined as:

- $R_{k,H}$: Transmission bit rate.
- $(E_b / N_0)_{k,H}$: BER parameter.

The offered traffic-load (in erl) of service-class k handoff calls is denoted as $a_{k,H}$ .We assume perfect power control. That is, at the BS, the received power from each service-class k call is the same and equal to P_k . Recall that in CDMA systems all MUs transmit within the same frequency band. Therefore, signals generated by MUs cause interference to each other. We distinguish the *intra-cell interference*, I_{intra} , caused by users of the reference cell and the *inter-cell interference*, I_{inter} , caused by users of the neighbouring cells. We also consider the existence of the thermal noise, P_N , which corresponds to the interference of an empty system.

The CAC in CDMA systems is performed by measuring the *noise rise*, *NR*, defined as the ratio of the total received power at the BS, I_{total} , to the thermal noise power, P_N [6]:

$$NR = \frac{I_{total}}{P_N} = \frac{I_{intra} + I_{inter} + P_N}{P_N} \quad (1)$$

When a new call arrives, the CAC estimates the noise rise and if it exceeds a maximum value, $NR_{\max}$, the new call is blocked and lost.

A service-class k call alternates between *active* (transmitting) and *passive* (silent) periods. This behavior is described by the *activity factor*, v_k , which represents the fraction of the call's *active* period over the entire service time ($0 < v_k \le 1$). Users that at a time instant occupy system resources are referred to as *active users*.

The *cell load, n*, is defined as the ratio of the received power from all *active* users (at the reference or neighbouring cells) to the total received power:

$$n = \frac{I_{intra} + I_{inter}}{I_{intra} + I_{inter} + P_N} \quad (2)$$

Hence from (1) and (2) we can derive the relation between the *noise rise* and the *cell load*:

$$NR = \frac{1}{1-n} \text{ and } n = \frac{NR-1}{NR} \quad (3)$$

We define the maximum value of the *cell load*, $n_{\max}$, as the *cell load* that corresponds to the maximum *noise rise*, $NR_{\max}$. A typical value in W-CDMA systems is $n_{\max} = 0.8$ and it can be considered as the shared system resource [6].

The *load factor*, $L_{k,N}$, given in (4) can be considered as the resource/bandwidth requirement of a service-class k new call:

$$L_{k,N} = \frac{(E_b / N_0)_{k,N} * R_{k,N}}{W + (E_b / N_0)_{k,N} * R_{k,N}} \quad (4)$$

By W we denote the chip rate of the W-CDMA carrier which is 3.84 Mcps.

The *cell load*, *n*, can be written (see (7) below) as the sum of the *intra-cell load*, n_{intra} (*cell load* that derives from the *active* users of the reference cell), and the *inter-cell load*, n_{inter} (*cell load* that derives from the *active* users of the neighbouring cells). They are defined in (5) and (6), respectively:

$$n_{intra} = \sum_{k=1}^{K} m_{k,N} L_{k,N} \quad (5)$$

ISBN: 978-1-61208-360-5

where m_k is the number of *active calls* among service-class k new calls.

$$n_{inter} = (1 - n_{max})\frac{I_{inter}}{P_N} \quad (6)$$

$$n = n_{intra} + n_{inter} \quad (7)$$

In this work, we adopt the following CAC condition at the BS in order to decide whether to accept a new service-class k call or not:

$$n + L_{k,N} \le n_{\max,N} \quad (8)$$

Similarly, the condition for the acceptance of a handoff service-class k call is:

$$n + L_{k,H} \le n_{\max,H} \quad (9)$$

where the derivation of $L_{k,H}$ is calculated similarly to (4).

IV. Local Blocking Probabilities

Due to the condition of (8), the probability that a new service-class k call is blocked when arriving at an instant with *intra-cell load*, n_{intra}, is called Local Blocking Probability (LBP) and can be calculated by [21]:

$$\beta_{k,N}(n_{intra}) = P(n_{intra} + n_{inter} + L_k > n_{max,N}) \quad (10)$$

In a similar way, we define the LBP for a handoff service-class k call:

$$\beta_{k,H}(n_{intra}) = P(n_{intra} + n_{inter} + L_k > n_{max,H}) \quad (11)$$

In order to calculate the LBP of (10) we can use (4)-(7). We notice that the only unknown parameter is the *inter-cell interference*, I_{inter}. Similarly to [21], we model I_{inter} as a lognormal random variable (with parameters μ_I and σ_I), that is independent of the *intra-cell interference*. Hence, the mean, $E[I_{inter}]$, and the variance, $Var[I_{inter}]$, of I_{inter} are calculated by (12) and (13):

$$E[I_{inter}] = e^{\mu_I + \frac{\sigma_I^2}{2}} \quad (12)$$

$$Var[I_{inter}] = (e^{\sigma_I^2} - 1)e^{2\mu_I + \sigma_I^2} \quad (13)$$

Consequently, because of (6), the *inter-cell load*, n_{inter}, will also be a lognormal random variable. Its mean, $E[n_{inter}]$, and variance, $Var[n_{inter}]$, are given by (14) and (15), respectively:

$$E[n_{inter}] = e^{\mu_n + \frac{\sigma_n^2}{2}} = \frac{1 - n_{max}}{P_N} E[I_{inter}] \quad (14)$$

$$Var[n_{inter}] = (e^{\sigma_n^2} - 1)e^{2\mu_n + \sigma_n^2} = (\frac{1 - n_{max}}{P_N})^2 Var[I_{inter}] \quad (15)$$

where μ_n and σ_n are the parameters of n_{inter}, which can be found by solving (16) and (17):

$$\mu_n = \ln(E[I_{inter}]) - \frac{\ln(1 + CV[I_{inter}]^2)}{2} + \ln(1 - n_{max}) - \ln(P_N) \quad (16)$$

$$\sigma_n = \sqrt{\ln(1 + CV[I_{inter}]^2)} \quad (17)$$

The coefficient of variation $CV[I_{inter}]$ is defined as:

$$CV[I_{inter}] = \frac{\sqrt{Var[I_{inter}]}}{E[I_{inter}]} \quad (18)$$

Note that (10) can be rewritten as:

$$1 - \beta_{k,N}(n_{intra}) = P(n_{inter} \le n_{max,N} - n_{intra} - L_{k,N}) \quad (19)$$

The Right Hand Side of (19), is the cumulative distribution function of n_{inter}. It is denoted by $P(n_{inter} \le n) = F_n(x)$ and can be calculated from:

$$F_n(x) = \frac{1}{2}[1 + erf(\frac{\ln x - \mu_n}{\sigma_n\sqrt{2}})] \quad (20)$$

where *erf*(•) is the well-known *error function*.

Hence, if we substitute $x = n_{\max,N} - n_{intra} - L_{k,N}$ into (20), from (19) we can calculate the LBP of new service-class k calls as follows:

$$\beta_{k,N}(n_{intra}) = \begin{cases} 1 - F_n(x), & x \ge 0 \\ 1, & x < 0 \end{cases} \quad (21)$$

Following a similar analysis, we can derive the LBP of handoff service-class k calls as follows:

$$\beta_{k,H}(n_{intra}) = \begin{cases} 1 - F_n(x), & x \ge 0 \\ 1, & x < 0 \end{cases} \quad (22)$$

V. State and Call Blocking Probabilities

A. State Probabilities

As stated before, in CDMA networks the *cell load* can be considered as a shared system resource and the *load factor* as the resource requirement of a call. Thus, we can use a modification of the Kaufman-Roberts recursion (K-R recursion) used for the determination of the link occupancy distribution in the EMLM [3], [4], for the calculation of *state probabilities* in CDTM systems. Below we present five steps needed for the modification.

The discretization of the *cell load*, *n*, and the *load factor*, $L_{k,N}$, is performed with the use of the *basic cell load unit*, *g*:

$$C = \frac{n_{max}}{g} \tag{23}$$

$$b_{k,N} = \text{round}(\frac{L_{k,N}}{g}) \tag{24}$$

where C is the system bandwidth capacity and $b_{k,N}$ is the bandwidth requirement of a new service-class *k* call.

We denote by *c* the total number of occupied b.u. at an instant and by *j* the total number of b.u. that would be occupied if all users were *active*. The parameter *j* at a given moment is considered as the system state.

We also denote by *q*(*j*) the probability of the state *j*. The *bandwidth occupancy*, $\Lambda(c \mid j)$, is defined as the conditional probability that *c* b.u. are occupied in state *j* and can be calculated from (25) recursively:

$$\begin{aligned}\Lambda(c \mid j) = &\sum_{k=1}^{K} P_{k,N}(j)[v_k \Lambda(c-b_{k,N} \mid j-b_{k,N}) + (1-v_k)\Lambda(c \mid j-b_{k,N})] \\ &+\sum_{k=1}^{K} P_{k,H}(j)[v_k \Lambda(c-b_{k,H} \mid j-b_{k,H}) + (1-v_k)\Lambda(c \mid j-b_{k,H})], \quad (25)\\ &\text{for } j = 1,...,j_{\max} \text{ and } c \le j\end{aligned}$$

where $j_{\max}$ is the max. system state, $\Lambda(0|0)$=1 and $\Lambda(c|j)$=0 for *c>j*.

In CDMA systems, due to the *inter-cell interference*, blocking of a service-class k call may occur at any state *j* with a probability $LB_{k,N}(j)$. This probability is given by summing over *c* the LBPs multiplied by the corresponding bandwidth occupancies:

$$LB_{k,N}(j) = \sum_{c=0}^{j} \beta_{k,N}(c)\Lambda(c \mid j) \tag{26}$$

The service-class *k bandwidth share* in state *j*, $P_{k,N}(j)$ and $P_{k,H}(j)$, can be derived from (27) and (28) for new and handoff calls, respectively.

$$P_{k,N}(j) = \frac{\alpha_{k,N}(1\text{-}LB_{k,N}(j\text{-}b_{k,N}))b_{k,N}q(j\text{-}b_{k,N})}{jq(j)} \tag{27}$$

$$P_{k,H}(j) = \frac{\alpha_{k,H}(1\text{-}LB_{k,H}(j\text{-}b_{k,H}))b_{k,H}q(j\text{-}b_{k,H})}{jq(j)} \tag{28}$$

The un-normalized *state probabilities* are given by extending the K-R recursion due to the presence of local blockings:

$$\begin{aligned}\hat{q}(j) = &\frac{1}{j}\sum_{k=1}^{K}\alpha_{k,H}(1\text{-}L_{k,H}(j\text{-}b_{k,H}))b_{k,H}\delta_k(j)\hat{q}(j\text{-}b_{k,H})+ \\ &\frac{1}{j}\sum_{k=1}^{K}\alpha_{kc,H}(1\text{-}L_{k,H}(j\text{-}b_{kc,H}))b_{kc,H}\delta_{kc}(j)\hat{q}(j\text{-}b_{kc,H})+ \quad (29)\\ &\frac{1}{j}\sum_{k=1}^{K}\alpha_{k,N}(1\text{-}L_{k,N}(j\text{-}b_{k,N}))b_{k,N}\hat{q}(j\text{-}b_{k,N}) \;, \\ &\text{for } j = 1,...,j_{\max}\end{aligned}$$

where $\hat{q}(0)$=1, $\hat{q}(j)$=0 for *j*<0 and the parameters $\delta_{k,H}(j), \delta_{kc,H}(j)$ are given by (30) and (31), respectively.

$$\delta_k(j) = \begin{cases} 1, & \text{when } 1 \le j \le C \text{ and } b_{kc} = 0 \\ 1, & \text{when } j \le J_k + b_k \text{ and } b_{kc} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{30}$$

$$\delta_{kc}(j) = \begin{cases} 1, \text{ when } (b_k \ge j > J_k + b_{kc}) \text{ and } (b_{kc} > 0) \\ 0, \text{ otherwise} \end{cases} \tag{31}$$

The threshold J_k is used for the selection of the bandwidth requirement of an arriving service-class *k* handoff call. In particular, if $j > J_k$, then the requested bandwidth is $b_{k,H}$; if $j \le J_k$, then the requested bandwidth is $b_{kc,H}$.

Finally, the normalized *state probabilities*, $q(j)$, are given by:

$$q(j) = \frac{\hat{q}(j)}{\sum_{j=0}^{j_{\max}} \hat{q}(j)} \tag{32}$$

B. Call Blocking Probabilities

The new-call blocking probabilities of service-class *k* can be calculated by adding all the *state probabilities* multiplied by the corresponding LBFs:

$$B_{k,N} = \sum_{j=0}^{j_{\max}} q(j)LB_{k,N}(j) \tag{33}$$

For the calculation of the *handoff-call* blocking probability, $B_{k,H}$, we must take into account the threshold J_k and thus incorporate the parameter $\gamma_k(j)$, defined as:

$$\gamma_k(j) = \begin{cases} 1, \text{ when } j \le J_k \\ 0, \text{ otherwise} \end{cases} \tag{34}$$

Hence, the *handoff-call* blocking probability of service-class k is given by:

$$B_{k,H} = \sum_{j=0}^{j_{\max}} q(j)\gamma_k(j) LB_{k,H}(j) \qquad (35)$$

VI. Performance Evaluation

In this section, we compare the analytical versus simulation results in respect of call blocking probabilities. The simulation language used is Simscript III [25]. We present analytical and simulation results for both types of calls, new and handoff. Simulation results are mean values of 10 runs with 95% confidence interval. The resultant reliability ranges of the simulation measurements are very small and, therefore, we present only mean values.

We evaluate two different service-classes with the following parameters:

a) $R_1 = 144\,Kbps, (E_b / N_0)_1 = 3\,dB$, and $v_1 = 0.67$.

b) $R_{2,1} = 384\,Kbps, R_{2,2} = 320\,Kbps$, $(E_b / N_0)_2 = 4\,dB$, $J_1 = 0.6$ and $v_2 = 1$.

We assume that the inter-cell interference is lognormally distributed with mean $E[I_{inter}] = 2 * E - 18\,mW$ and coefficient of variation $CV[I_{inter}] = 1$. The thermal noise power density is $-174\,dBm / Hz$. For discretization we use $g = 0.001$. The following cell load thresholds are considered: $n_H = n_{\max} = 0.8$ and $n_N = 0.75$, for new and handoff calls, respectively. We generate traffic load according to the Table I. That is, in the case of the 1st service-class the traffic-load point 1 corresponds to $a_{1,N} = 1.0$ erl and $a_{1,H} = 0.1$ erl for the new and handoff calls, respectively.

In Figs. 1 and 2, we present the analytical and simulation call blocking probabilities for the 1st and the 2nd service-class, respectively. We observe that the accuracy of the proposed model is very good, since the analytical results are very close to simulation results in all cases. We also observe that by using different cell load thresholds for new and handoff calls ($n_N > n_H$), we achieve lower blocking probabilities for the handoff calls.

TABLE I. OFFERED TRAFFIC (ERL).

	1	2	3	4	5	6
$a_{1,N}$	1.0	1.25	1.5	1.75	2.0	2.25
$a_{1,H}$	0.1	0.2	0.3	0.4	0.5	0.6
$a_{2,N}$	0.2	0.3	0.4	0.5	0.6	0.7
$a_{2,H}$	0.05	0.1	0.15	0.2	0.25	0.3

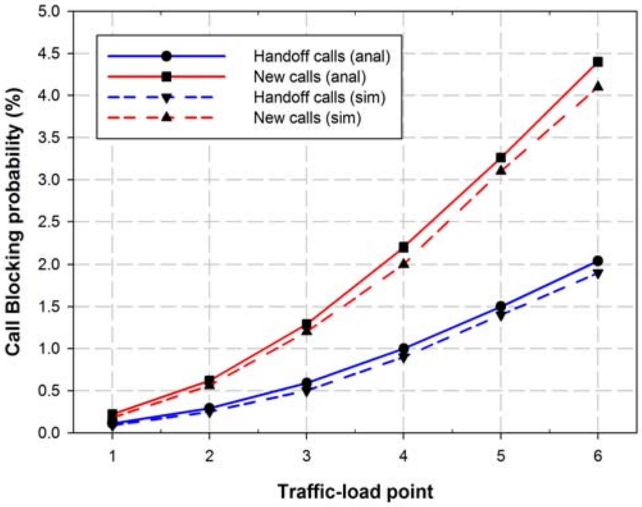


Figure 1. Call blocking probabilities for new and handoff calls vs offered traffic-load (1st service-class).

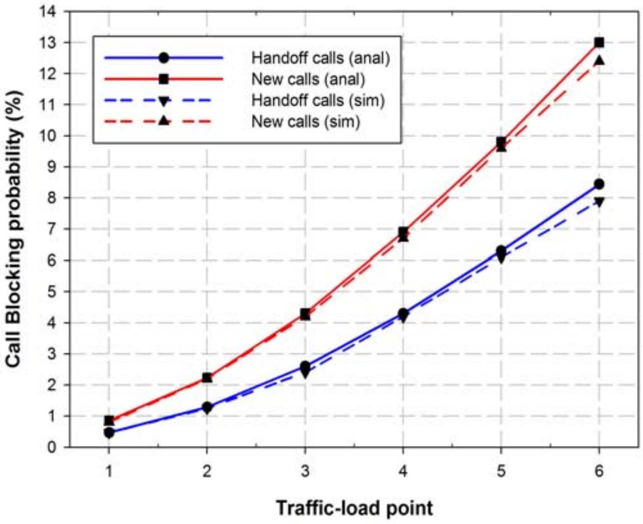


Figure 2. Call blocking probabilities for new and handoff calls vs offered traffic-load (2nd service-class).

VII. Conclusions and Future Work

In this paper, we have presented a new teletraffic model for the analysis of handoff traffic in cellular CDMA systems. The proposed approach is based on a single-threshold model that allows the handoff call request less bandwidth when the system is overloaded. In that case, the handoff-call blocking probability can be reduced. We have performed simulation studies, which show that the accuracy of our proposed analytical model is very satisfactory. As a future work, we will study the impact of multiple thresholds on the blocking probabilities of handoff calls. Also, we will incorporate a finite number of traffic sources into our model.

REFERENCES

[1] "Intelligent base stations" (white paper), Nokia Siemens Networks, 2012.

[2] J. Kaufman, "Blocking in a shared resource environment," IEEE Transactions on Communications, vol. 29, no. 10, October 1981, pp. 1474–1481.

[3] J. Roberts, "A service system with heterogeneous user requirements," in: G. Pujolle (Ed.), Performance of Data Communications Systems and Their Applications, North Holland, Amsterdam, 1981, pp. 423-431.

[4] J. Kaufman, "Blocking with retrials in a completely shared resource environment," Performance Evaluation, vol. 15, no. 2, June 1992, pp. 99-113.

[5] M. Sobieraj, M. Stasiak, J. Weissenberg, and P. Zwierzykowski, "Analytical model of the single threshold mechanism with hysteresis for multi-service networks," IEICE Transactions on Communications, vol. 95-B, no. 1, 2012, pp. 120–132.

[6] D. Staehle and A. Mäder, "An analytic approximation of the uplink capacity in a UMTS network with heterogeneous traffic," Proc. 18th International Teletraffic Congress (ITC), Berlin, August 31 – September 5, 2003.

[7] G. Kallos, V. Vassilakis, I. Moscholios, and M. Logothetis, "Performance modelling of W-CDMA networks supporting elastic and adaptive traffic," Proc. Fourth International Working Conference on Performance Modelling and Evalautaion of Heterogeneous Networks (HET-NETs 2006), Ilkley, U.K., Sept. 2006.

[8] V. Vassilakis, G. Kallos, I. Moscholios, and M. Logothetis, "The wireless Engset multi-rate loss model for the call-level analysis of W-CDMA networks," Proc. IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2007), Athens, Greece, Sept. 2007.

[9] M. Stasiak, A. Wisniewski, P. Zwierzykowski, and M. Glabowski, "Blocking probability calculation for cellular systems with WCDMA radio interface servicing PCT1 and PCT2 multirate traffic," IEICE Transactions on Communications, vol. E92-B, no. 4, April 2009, pp. 1156-1165.

[10] V. Vassilakis, G. Kallos, I. Moscholios, and M. Logothetis, "Evaluation of a call admission control scheme in W-CDMA cellular networks," Proc. IEEE Sixth International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP'08), Graz, Austria, July 2008.

[11] D. Parniewicz, M. Stasiak, and P. Zwierzykowski, "Analytical model of the multi-service cellular network servicing multicast connections," Telecommunication Systems, Springer, vol. 52, no. 2, February 2011, pp. 1091-1100.

[12] V. Vassilakis, G. Kallos, I. Moscholios, and M. Logothetis, "Call-level analysis of W-CDMA networks supporting elastic services of finite population," Proc. IEEE International Conference on Communications (ICC 2008), Beijing, China, May 19-23, 2008.

[13] G. Kallos, V. Vassilakis, and M. Logothetis, "Call-level performance analysis of a W-CDMA cell with finite population and interference cancellation," European Transactions on Telecommunications, vol. 22, January 2011, pp. 25–30.

[14] I. Moscholios, G. Kallos, V. Vassilakis, M. Logothetis, and M. Koukias, "Congestion probabilities in W-CDMA networks supporting calls of finite sources," Proc. Seventh International Conference on Performance and Security Modelling & Evaluation of Cooperative Heterogenious Networks (HET-NETs 2013), Ilkley, West Yorkshire, U.K, November 11-13, 2013.

[15] I. Moscholios, G. Kallos, M. Katsiva, V. Vassilakis, and M. Logothetis, "Call blocking probabilities in a W-CDMA cell with interference cancellation and bandwidth reservation," Proc. IEICE Information and Communication Technology Forum (ICTF), Poznan, Poland, May 28-30, 2014.

[16] M. Ivanovich and P. Fitzpatrick, "An accurate performance approximation for beyond 3G wireless broadband systems with QoS," IEEE Trans. on Veh. Tech., vol. 62, no.5, June 2013, pp. 2230-2238.

[17] G. Kallos, V. Vassilakis, and M. Logothetis, "Call blocking probabilities in a W-CDMA cell with fixed number of channels and finite number of traffic sources," Proc. IEEE Sixth International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP 2008), Graz, Austria, 23-25 July 2008.

[18] I. Moscholios, M. Katsiva, G. Kallos, V. Vassilakis, and M. Logothetis, "Equalization of congestion probabilities in a W-CDMA cell supporting calls of finite sources with interference cancellation," Proc. IEEE/IET 9th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP 2014), Manchester, U.K., 23-25 July 2014.

[19] M. Stasiak, P. Zwierzykowski, and D. Parniewicz, "Modelling of the WCDMA interface in the UMTS network with soft handoff mechanism," Proc. IEEE Global Communications Conference, (GLOBECOM 2009), Honolulu, Hawaii, USA, 2009.

[20] V. Vassilakis and M. Logothetis, "The wireless Engset multirate loss model for the handoff traffic analysis in W-CDMA networks," Proc. IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2008), Cannes, France, August 31 – September 4, 2008.

[21] V. Vassilakis, G. Kallos, I. Moscholios, and M. Logothetis, "On the handoff-call blocking probability calculation in W-CDMA cellular networks," Proc. IARIA 4th Advanced International Conference on Telecommunications (AICT 2008), Athens, Greece, June 8-13, 2008.

[22] V. Vassilakis, G. Kallos, I. Moscholios, and M. Logothetis, "On call admission control in W-CDMA networks supporting handoff traffic," Ubiquitous Computing and Communication Journal - Special issue on Communication Systems, Networks and Digital Signal Processing, January 2009.

[23] A. Mäder and D. Staehle, "Analytic modeling of the WCDMA downlink capacity in multi-service environments," Proc. 16th International Teletraffic Congress (ITC) Specialist Seminar, Antwerp, Belgium, August 31 – September 2, 2004.

[24] I. Daskalopoulos, V. Vassilakis, and M. Logothetis, "Thorough analysis of downlink capacity in a WCDMA cell," Proc. First International Conference on Mobile Lightweight Wireless Systems (MOBILIGHT 2009), Athens, Greece, May 18-20, 2009.

[25] Simscript III, http://www.simscript.com [May 2014].

Design and Implementation of Simulation Engine for Very High-Rate Communication over Power Grid

Sungsoo Choi and Hui-Myoung Oh
Power Telecommunication Research Center
Korea Electrotechnology Research Institute (KERI)
Ansan, S. Korea
{sschoi, hmoh}@keri.re.kr

***Abstract*—This paper discusses the design of an engine to simulate a Very high-rate Power Line Communication (VPLC) infrastructure that handles up to a rate of 400 Mbps. The simulation engine complies with the ISO/IEC 12139-1 standard for power line communication protocols. The engine can be used as one of the access network elements deployed in the advanced metering infrastructure in a smart grid. We first propose a feasible system model of VPLC in line with the design goals with pre-simulations, i.e., an event-driven simulation and a timing-driven simulation. Next, we design a semiconductor Intellectual Property (IP) core, focusing on the implementation of two main functional IP cores of the entire system: a fast Fourier transform for modulation/demodulation and a low-density parity-check encoder/decoder for forward-error correction. Finally, the designed VPLC based on the main functional IP cores is implemented on an available FPGA chipset, targeting the Virtex-6 xc6vlx240T.**

Keywords—very high-rate power line communications; advanced metering infrastructure; smart grid; FPGA; ISO/IEC 12139-1.

I. Introduction

The Smart Grid (SG) plays a major role in achieving a low carbon footprint and is therefore a key component of the sustainable energy infrastructure. Since 2010, to achieve the vision of “Low carbon, green growth,” an SG test-bed has been built in Jeju island, Korea. Several power IT projects have been initiated integrating electric power technology as well as Information and Communication Technology (ICT) in five implementation areas, namely the smart power grid, smart consumer, smart transportation, smart renewables, and smart electricity service [1].

These implementation areas have completely different electrical environments and use different types PLC channels. Therefore, to meet various technical requirements for different network communication scenarios, several protocols and advanced modulation techniques are employed by Power Line Communication (PLC) systems in the SG. In addition, both the broadband spectrum from 1.8 MHz to 30 MHz (or 205 MHz) as well as the narrowband spectrum from 3 kHz to 500 kHz [2,3] are used for PLC. Advanced Metering Infrastructure (AMI), a typical SG application, has been successfully implemented using broadband high-rate communication at 24 Mbps over the power grid in compliance with the ISO/IEC 12139-1 standard, which is based on the Korea Standard (KS) 4600-1 [3,4].

In Section II, we briefly introduce the specifications of the engine designed for Very high-rate Power Line Communication (VPLC) that is compliant with ISO/IEC 12139-1, by introducing a static simulation with an event-driven operation and a dynamic simulation with a timing operation. In Section III, we explain the design and implementation of two Intellectual Property (IP) cores, namely MOD (for the functions of modulation/ demodulation) and FEC (for the function of error correction), at the system level with a cycle-based design approach, which is based on the design methodology at a clock-based register transfer level. Appropriate hardware-efficient algorithms are selected to design the hardware architectures, taking into account resource constraints of timing and area. All the design steps are carried out using both floating-point and fixed-point operations to ensure design consistency from the system level to the architecture level. Actually, both floating-point and fixed-point designs are useful to obtain a test bench for verifying the designed digital logic block at the 32-bit level, compared with the results of the dynamic simulation. Section IV explains how for a hardware gate level, specific functional blocks are edited by a Hardware Description Language (HDL), which enables the synthesis of logic gates in a 32-bit process. The IP cores are designed to support high-speed digital signal processing using a pipelining technique, parallelizing technique, and retiming technique to meet the system requirements. Finally, the main IP core circuit of VPLC is realized using a Field-Programmable Gate Array (FPGA) and is presented on the prototyping board. Conclusion is presented in Section V.

II. System Design Issues for VPLC

A. IP Design Issues for VPLC

VPLC is proposed to support utility applications in the implementation areas of the smart consumer and the smart transportation. To achieve this, we design a system with the following basic design goals:

- Supporting maximum rates of 200 Mbps and 400 Mbps in a dual transmission rate mode, in the frequency bandwidth less than 30 MHz and less than 80 MHz

- Adopting a multicarrier modulation technique for VPLC, such as Discrete Multi-Tone (DMT) modulation or Orthogonal Frequency-Division Multiplexing (OFDM)
- Adopting a Forward Error Correction technique (FEC) of a Low Density Parity Check (LDPC) to improve the link margin by 8.8 to 9.4 dB at BER = 10^{-5}
- Supporting coexistence of other systems compliant with ISO/IEC 12139-1 (or KS 4600-1) [3, 4].

B. *Specifications of VPLC engine*

To set the specifications of the VPLC engine, we first extract the main parameters from the allowable spectral efficiency from the typical multicarrier transmission system. After estimating the main parameters, we consider ways of improve the system performance to meet the system design goals. The formula to calculate the multicarrier transmission rate can be written as follows:

$$Bps = B_w \cdot M \cdot \frac{B_u}{B_w} \cdot \frac{T_{FFT}}{T_{FFT} + T_{CP}} \cdot C_r \qquad (1)$$

where B_w is the entire signal bandwidth, B_u is the usable signal bandwidth, M is the number of bits in each subcarrier, T_{FFT} is the signal time to process a Fast Fourier Transform (FFT), T_{CP} is a cyclic prefix time, and C_r is the code rate for error correction. For designing a feasible system, the required overhead has to be adjusted, which can be decided by the factor of a guard band (B_u/B_w), a guard interval (GI) time ($T_{FFT}/(T_{FFT} + T_{CP})$), and a code rate ($C_r$). When considering a low frequency band of 0–28 MHz, the available bandwidth ratio becomes 0.8667. For convenience, we set the number of subcarriers to 2048 and the frequency band to 60 MHz in a high frequency band to reach a goal of greater than 400 Mbps. As shown in Table I, we need to set the main parameters as at least 1024-QAM modulation in the subcarrier for OFDM/DMT, the code rate greater than 7/8, and the guard interval time less than 2.13 µs.

By setting these parameters, we can achieve the required received *SNR* greater than 34 dB to meet the 1% packet error rate requirement for a typical multicarrier transceiver with a simple receiver structure in an ideal additive white Gaussian noise channel without any reflection phenomenon, as shown in Fig. 1.

TABLE I. MAIN PARAMETERS FOR MULTICARRIER TRANSMISSION.

CR	**Modulation in a Sub-Carrier for OFDM /DMT**	**Number of Coded Bits per Symbol**	**Number of Bits per Symbol**	**Data Rate (Mbps)**		**Spectral Efficiency (bps/Hz)**	
				GI = 4.27 µs	GI = 2.13 µs	GI = 4.27 µs	GI = 2.13 µs
1/2	256 QAM	13824	6912	180	190.5	3	3.17
3/4	256 QAM	13824	10368	270	285.8	4.5	4.76
5/6	256 QAM	13824	11520	300	317.6	5	5.29
7/8	256 QAM	13824	12096	315	333.5	5.25	5.55
1/2	1024 QAM	17280	8640	225	238.2	3.75	3.97
3/4	1024 QAM	17280	12960	337.5	357.3	5.62	5.95
5/6	1024 QAM	17280	14400	375	397.0	6.25	6.61
7/8	1024 QAM	17280	15120	393.75	416.9	6.56	6.94

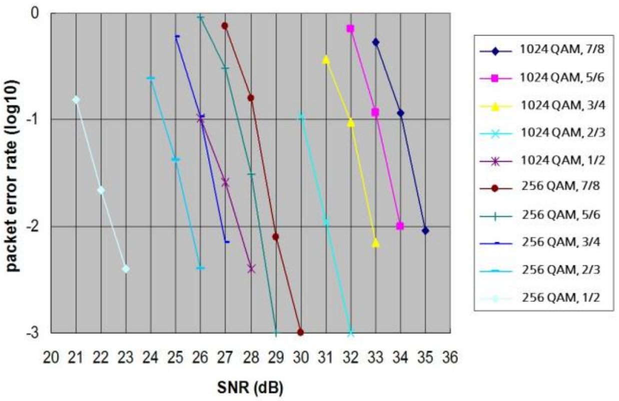


Figure 1. Performance of a Typical Multicarrier Transceiver.

Even under the real channel environment conditions of a power grid, it is possible to ensure more margins to meet the basic performance requirements. We consider additive techniques such as a data frame header check sequence [4] for minimizing the SNR, and a channel encoder [7,8] to protect the data frame header such as the LDPC. For designing an LDPC with a variable code rate of 1/2, 2/3, 3/4, 5/6, and 7/8, the system complexity can be effectively reduced by using the circular shift technique, considering a size-limited unit matrix of 24 by 24, to operate the large matrix of the LDPC.

TABLE II. SPECIFICATION OF VPLC.

Features	**DMT Symbol variables**		
	Preamble Frame	**Control Frame**	**Long Symbol Frame, Data Frame Header, Data Frame**
Bandwidth	25 MHz	25 MHz	75 MHz
Sampling Frequency	50 MHz	50 MHz	200 MHz
Tone Space	97.65625 kHz	97.65625 kHz	48.828125 kHz
IFFT Space	512 Samples	512 Samples	4096 Samples
Prefix Space	0 Samples	128 Samples	448 Samples
Roll-off Space	16 Samples	16 Samples	-
Symbol Duration	512 Samples	624 Samples	4544 Samples
FFT Period	10.24 µs	10.24 µs	20.48 µs
Symbol Length	10.24 µs	12.48 µs	22.72
Tone(or sub channel) Modulation	16 PSK	DBPSK	BPSK, QPSK, 16 QAM, 64 QAM, 256 QAM, 1024 QAM

In addition, to improve the system reliability, we consider a bit-loading technique to adjust variable modulation selection and variable channel code-rate selection in the subcarriers even at the cost of the transmission rate. Further, to overcome the performance degradation due to an impulsive noise on the power grid, we adopt a diversity technique, i.e. repeatedly transmitting a signal block in the time domain and using subcarrier utilization in the frequency domain. In other words, to ensure the compliance of the given deployed transceiver to

ISBN: 978-1-61208-360-5

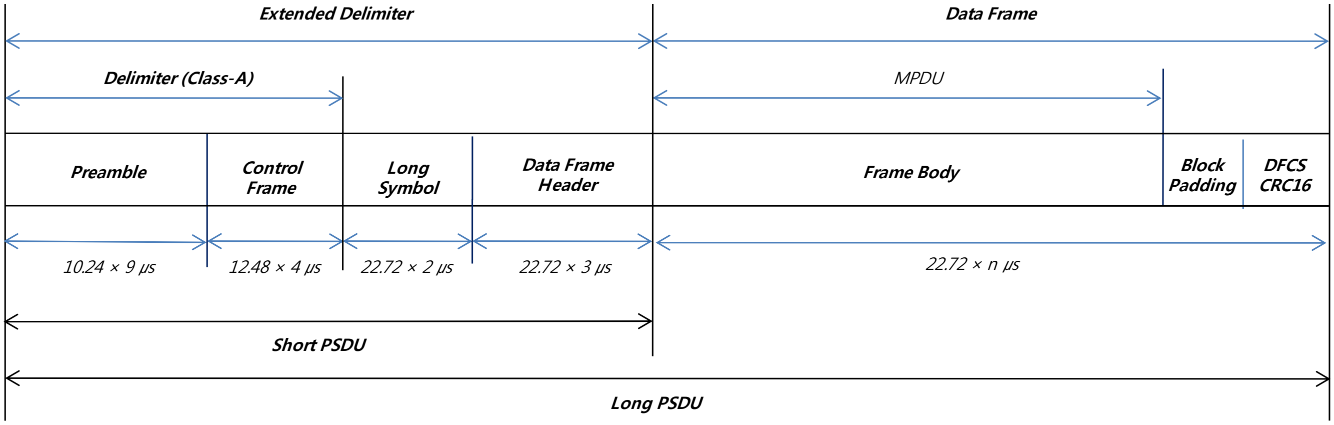


Figure 2. Transmitted VPLC Packet Service Data Unit Signal.

the ISO/IEC 12139-1 standard, we use an ISO/IEC 12139-1 based check sequence in the control frame, diversity mapping, Differential Phase Shift Keying (DPSK) modulation in subcarriers, forward error correction of RS(5,3), and 512-point FFT. Table II lists the specifications of VPLC. As shown in Fig. 2, the transmitted VPLC Packet Service Data Unit (PSDU) signal consists of a preamble frame, a control frame, and the rest of frames. The Delimiter signal has the preamble frame and the control frame. The Extended Delimiter has additional frames of the Long Symbol frame and the Data Frame Header. The Long Symbol consists of 2 DMT symbols and the Data Frame Header consists of 3 DMT symbols. The Data Frame consists of *n* DMT symbols. Each DMT symbol has 1,536 subcarriers in the frequency band of 75 MHz. The subcarrier bandwidth is 48.828125 kHz, and each DMT symbol length is 22.72 μs, because of the addition of a cyclic prefix with 2.24 μs for minimizing multipath channel effects. The DMT symbol sets to the identical field of the Delimiter signal according to the Class-A version of ISO/IEC 12139-1 for supporting the function of coexistence. The structure of the PSDU can be divided into the long PSDU and the short PSDU whether the Data Frame is included or not.

C. System Simulation for VPLC over Power Grid

1) Event-Driven (ED) Simulation

According to the proposed specification of VPLC, we first develop the ED simulator by using C++ to check the feasibility. The ED simulator can be independently operated by appropriate selection of the inter-frame types in PSDU. The transmit data, generated from a Medium Access Control (MAC) layer, pass through the first functional block viz. the scrambler that removes the unique pattern of the transmit data. In the following FEC block of LDPC, the transmit data is encoded to prevent data redundancy. At the Mapper block, the encoded data stream is converted to the bit-to-symbol data for the purpose of allotting it into each subcarrier. Next, the sequence of data is added to process the modulation step by using Inverse Fast Fourier Transform (IFFT) [6], giving an orthogonal property between subcarriers. The rest of the transmit signals except a Preamble signal are added by a Cyclic Prefix sequence to minimize the effect of an Inter-Symbol Interference (ISI) in the communication link channel. At the receiver side, the receive signal can be recognized by adding the Carrier Sensing block for detecting the signal frame, and by synchronizing the symbol to the starting point of the DMT symbol. Through the Windowing block, the size of the received signal sequence is reformatted to the length of FFT, and then it is demodulated through the FFT block. First of all, using the Long symbol, the receiver estimates the status of the given channel and compensates a timing error caused by the difference in the clock frequencies of the digital-to-analog converter and the analog-to-digital converter. The Demapper computes the log-likelihood ratio in each binary data. The LDPC decoder corrects the channel errors by using the parity information. Finally, in the descrambler block, the transmitted signal can be recovered. In this simulation, we use a conventional power line channel model such as a wireless Rayleigh channel model [5] as follows:

$$H(f) = \sum_{i=1}^{N} g_i \, e^{-\left(a_0 + a_1 f^k\right) d_i} \, e^{-j 2\pi f \left(d_i / v_p\right)} \qquad (2)$$

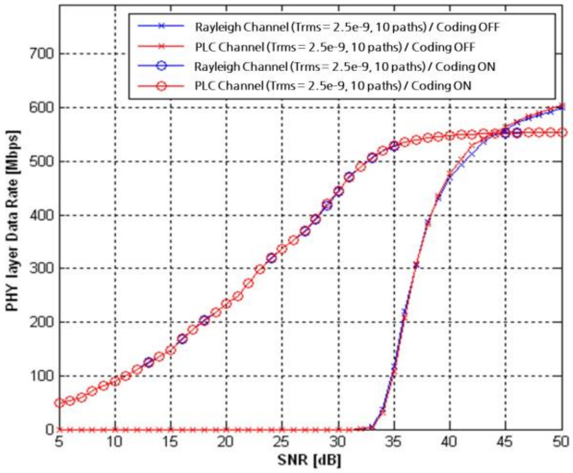


Figure 3. Simulation Result based on ED simulator.

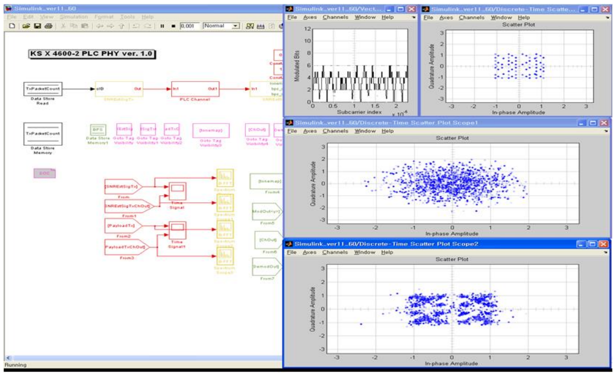

Simulink_ver11_60
File Edit View Simulation Format Tools Help
0.001 Normal
KS X 4600-2 PLC PHY ver. 1.0
TxPacketCount
Data Store Read
PLC Channel
Data Store Memory
Time Signal
Time Signal1
Simulink_ver11_60/Vect...
Simulink_ver11_60/Discrete-Time Scatte...
Simulink_ver11_60/Discrete-Time Scatter Plot Scope1
Simulink_ver11_60/Discrete-Time Scatter Plot Scope2
File Axes Channels Window Help
Modulated Bits
Subcarrier index
Scatter Plot
Quadrature Amplitude
In-phase Amplitude
Running

Figure 4. TD Simulator and Simulation Result.

where *i* is the number of paths, g_i is a weighting factor, *k* is an attenuation factor of exponent, a_0 and a_1 are attenuation parameters, and d_1 is the length of the path [5]. Fig. 3 shows communication performance of the proposed VPLC in the ED simulation for the Rayleigh channel model in (2) with the parameter of time = 2.5 e^{-9} s and 10 paths. It meets the minimum communication requirement of the signal-to-noise ratio (SNR) of about 28 dB to guarantee the rate of 400 Mbps. Without any FEC, to guarantee the rate of 400 Mbps, an SNR greater than approximately 37 dB is required.

2) *Timing-Driven (TD) Simulation*

The ED simulation is for a static design approach, whereas the TD simulation is for a more specific timing design approach, which operates in a symbol time basis. The GUI environment of the TD simulator is developed by using the S/W tool of the Simulink and the Matlab of MathWorks. The sample values can be analyzed and verified by both floating-point and fixed-point numerals, which support functional features required to design the architecture for controlling each parameter of the structural elements in the VPLC system. Besides, the status of the current signal stream can be checked by plotting the transmission symbols in both time and frequency domains. Fig. 4 shows the symbol-by-symbol transmission from the transmitter to receiver and a symbol-based scattering plot. The TD simulator developed was found to operate satisfactorily in accordance with the specifications of VPLC.

III. IP Design Issues for VPLC

From the simulator introduced in previous section, reference vectors called a test bench can be extracted for basically designing and testing the specific logic architecture of IP in the chipset. In this section, we briefly study the IP design issues by introducing a couple of important IP engines, a MOD and a FEC for the proposed VPLC system.

A. *Designing Signal Process of MOD*

The MOD is one of core engines for signal processing and integrating the VPLC system. The engine supports the modulation/demodulation function in the proposed VPLC system, adopting an Orthogonal Frequency Division Multiplex (OFDM) technique. Functionally, the MOD maps binary data into OFDM signals, based on the modulation coefficients, by computing the 4096-point FFT. However, the complex 4096-point FFT may cause a bottleneck related its processing in a given clock time period. Therefore, a special architecture needs to be considered to effectively resolve this design problem related to timing and sizing. For an OFDM transmitter, the MOD can be processed by the Inverse FFT (IFFT) and it can be structurally designed just by hardwiring and switching the input and output ports of the FFT. The typical formula of the FFT is as follows [6]:

$$X\left[k\right]=\sum_{n=0}^{N-1}x[n]W_N^{nk} \qquad for\ \ k=0,1,...,N-1$$

$$W_N^{nk}=e^{-j\frac{2\pi}{N}nk} \qquad twiddle\ \ factor \qquad (3).$$

We adopt a Decimation-In-Frequency (DIF) FFT algorithm with Radix 4 (R4), which allows an easy design to have a semi-systolic parallel structure. The output sequence of the Radix-4 FFT is decimated as

$$\begin{aligned}
X[4r]&=\sum_{n=0}^{(N/4)-1}g_1[n]\cdot W_{N/4}^{nr}\\
X[4r+1]&=W_N^{n}\sum_{n=0}^{(N/4)-1}g_2[n]\cdot W_{N/4}^{nr}\\
X[4r+2]&=W_N^{2n}\sum_{n=0}^{(N/4)-1}g_3[n]\cdot W_{N/4}^{nr}\\
X[4r+3]&=W_N^{3n}\sum_{n=0}^{(N/4)-1}g_4[n]\cdot W_{N/4}^{nr}
\end{aligned} \qquad (4)$$

where

$$\begin{aligned}
g_1[n]&=x[n]+x[n+(N/4)]+x[n+(N/2)]+x[n+(3N/4)]\\
g_2[n]&=x[n]-jx[n+(N/4)]-x[n+(N/2)]+jx[n+(3N/4)]\\
g_3[n]&=x[n]-x[n+(N/4)]+x[n+(N/2)]-x[n+(3N/4)]\\
g_4[n]&=x[n]+jx[n+(N/4)]-x[n+(N/2)]-jx[n+(3N/4)]
\end{aligned}.$$

Figs. 5 and 6 show the proposed architecture of the 4096-point FFT with parallelizing and pipelining. When designing the architecture of the R4 FFT algorithm with an internal clock speed that takes into account the critical path delay by adopting the design approach of a Single-path Delay Commutator (SDC) with a pipelined architecture. This architecture has the advantage of improving the utilization of the Butterfly module used to compute the complex multiplication performed repeatedly in each stage of the 4096-FFT. Each element of the Butterfly module consists of a 4-point DFT process.

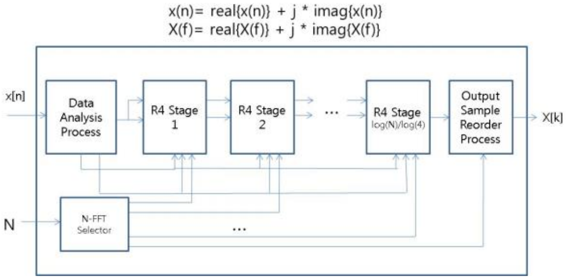

x(n)= real{x(n)} + j * imag{x(n)}
X(f)= real{X(f)} + j * imag{X(f)}
x[n]
Data Analysis Process
R4 Stage 1
R4 Stage 2
...
R4 Stage log(N)/log(4)
Output Sample Reorder Process
X[k]
N
N-FFT Selector
...

Figure 5. Architecture of the Proposed N-FFT (N = 4096).

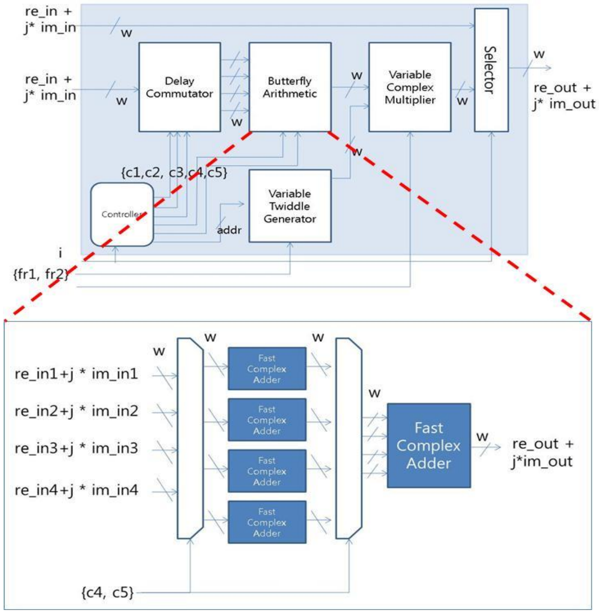


Figure 6. Architecture of the Basic R4 Butterfly Module in *i* th Stage in FFT.

Furthermore, to reduce the buffer size of the SDC, we logically design a RAM-type SDC to minimize quantization errors without any changes to the hardware architecture. By varying the magnitude of the input signal in the 4096-point FFT, the number of fixed point numerals can be automatically controlled. These are done in the front element of the data analysis process, which can be simply designed as a look-up table. In each stage, we consider controlling a variable twiddle-factor generator and complex multiplication with a fast complex adder. To reduce the complexity of the 4096-point FFT chip-level implementation, the basic functional blocks are shared. By sharing, the utilization factor can be increased up to 75% for the computation of complex multiplications. We obtain the result of the physical timing simulation and a Netlist of synthesized logic gate for targeting the Xilinx Vertex-6 chip. For verification of the design, we perform a harness test between the input/output signal results of the upper design layer with floating point values and that of the lower design layer with fixed point values. Furthermore, the input/output signals with binary data from the result of the HDL timing simulation are also verified with the harness test using the test bench of the outputs from the TD simulator.

B. *Designing Signal Process of FEC*

We design a FEC core engine by using a LDPC, which has features of the parity-check matrix with code rates of 1/2, 2/3, 3/4, (4/5), 5/6, (6/7), 7/8 and with code words of 576, 864, 1152, (1440), 1728, (2016), 2304. When designing the FEC, the LDPC uses a very large parity-check matrix for improving the decoding performance. In addition, it has different code rates in the proposed specification for VPLC. It means that the structure of the LDPC uses an individual parity-check matrix. Thus, it severely causes the problem of complexity in chip implementation. To reduce the complexity, the concept of sharing the function of a base matrix effectively is used. It can be operated adaptively according to the constant variation informative matrix in the parity-check matrix. We adopt the Richardson's algorithm for designing the architecture of the parity-check matrix and a structure of Quasi-Cycle (QC) for the submatrix [7,8]. Each submatrix uses a unit matrix with the size of G × G in accordance with the values of a circular cyclic shift in the parity-check matrix. It can be expressed as

$$\mathbf{I}^{s} = \left(a_{ij}\right), \quad a_{ij} = \begin{cases} 1, & if \ \mathrm{mod}_{G}(i+s) = j \\ 0, & otherwise \end{cases} \quad . \tag{5}$$

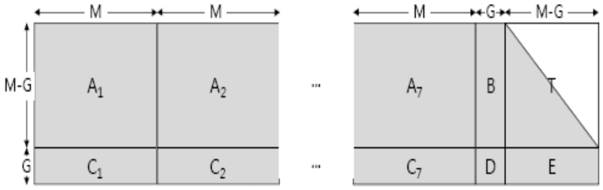


Figure 7. Structure of Parity-Check Matrix of LDPC for the FEC IP.

For the structure of the parity-check matrix, a redundancy part has a fixed length of M, and it can be changed from M to 7M, as shown in Fig. 7. In the part of redundancy, the parity vector p_1^T and p_2^T can be derived as follows:

$$\begin{aligned} \mathbf{p}_1^{\mathbf{T}} &= -\varphi^{-1}\left(-\mathbf{E}\mathbf{T}^{-1}\mathbf{A} + \mathbf{C}\right)\mathbf{u}^{\mathbf{T}} \\ \mathbf{p}_2^{\mathbf{T}} &= -\mathbf{T}^{-1}\left(\mathbf{A}\mathbf{u}^{\mathbf{T}} + \mathbf{B}\mathbf{p}_1^{\mathbf{T}}\right) \\ \varphi &= -\mathbf{E}\mathbf{T}^{-1}\mathbf{B} + \mathbf{D} \quad . \end{aligned} \tag{6}$$

Fig. 8 shows the architecture of the LDPC encoder, composed of the 24 × 24 submatrix (i.e., G = 24), with the length of information bits of 24 × 12 × n (where n = 1–7), and the final code words of 24 × 12 × (n + 1). For encoder design, we adopt the pipelining process technique with four stages, supporting the 24 clocks per each stage and the latency of 96 clocks. As shown in Fig. 8, we use circular shift registers for storing and dividing the input information bits according to code rates at the first step in the encoder. We devise on an arithmetic process for computing the sequential multiplication of the parity-check matrix and the information bits, and the parity vector arithmetic process (6) at the second step and the third step, respectively. In the last step, we obtain the output combining the parity vector. In other words, the decoder is designed for a variable iteration number (n) of the pipelining architecture with 2(n + 1) + 2 stages, 288 clocks, and 288(2n + 4) latency clocks. Besides, we set a stop option of iteration as necessary. The decoder consists of a PreDecoder with two stages, followed by a IterationDecoder with two stages and a ParityChecker with two stages, and, lastly, a Terminator with two stages, as shown in Fig. 9. Thus, according to the channel environment, a proper code rate from the designed LDPC can be obtained by controlling the information bits.

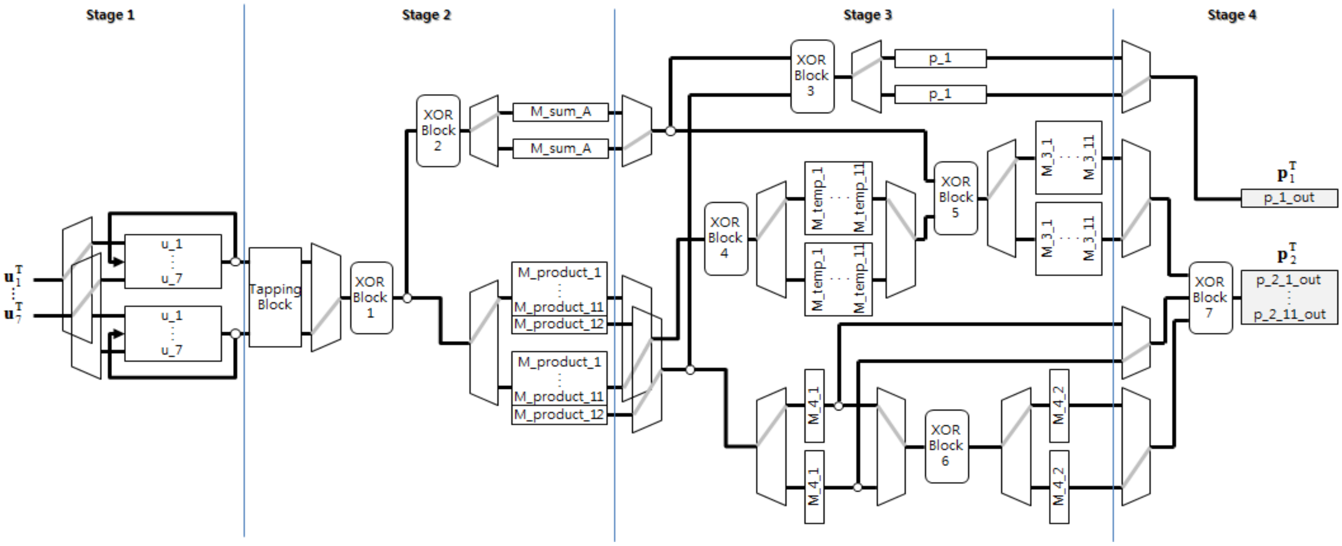


Figure 8. Architecture of the LDPC Encoder for the FEC IP.

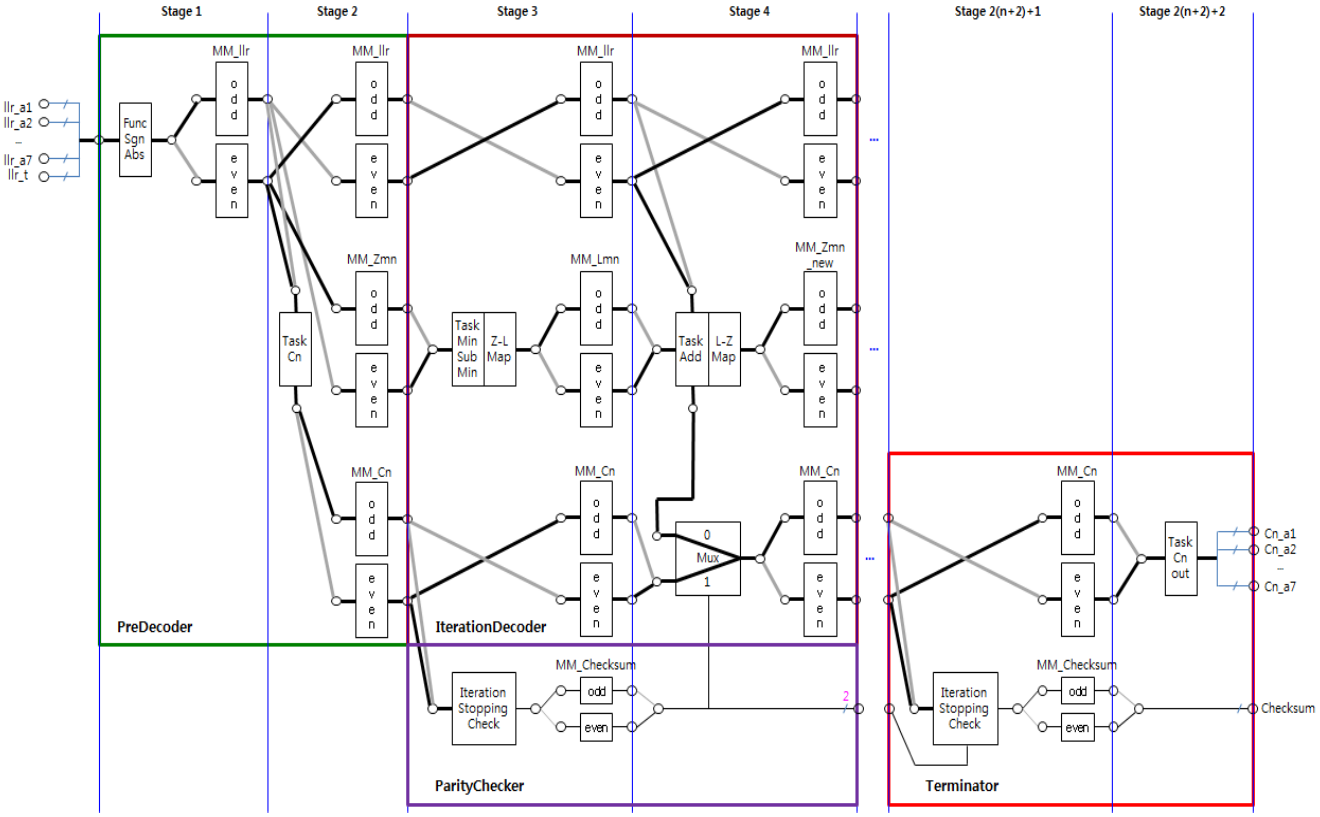


Figure 9. Architecture of the LDPC Decoder for the FEC IP.

IV. IMPLEMENTATION OF TEST PLATFORM

After the back-end design of the VHLC has been completed, the physically synthesized logic, combined with the IP of MCU, a 32-bit SE3208 microprocessor, is mounted on the Xilinx FPGA, targeting Virtex-6 xc6vlx240T. The test platform is developed and demonstrated to transfer the data with a moving picture as a real operation, as shown in Fig. 10. It has an operational function to monitor streaming data and directly control system configurations on the test board.

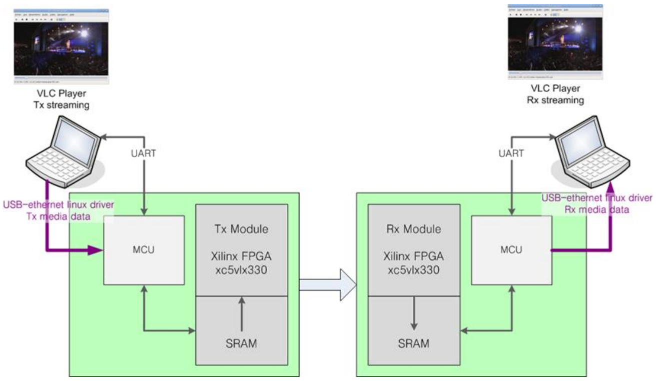


Figure 10. Test Platform for the VPLC Physical Layer.

V. Conclusion

In this study, we proposed a VPLC and demonstrated its feasibility of operation at a rate of more than 400 Mbps in the physical layer of the power grid environment. Further, we designed and implemented two core engines, namely the MOD and the FEC, necessary for simulating the VPLC system. We proposed and adopted the appropriate design algorithms and the design architecture to meet system specifications and requirements so that sharing main functional modules will reduce the system complexity and parallelize it effectively. All the designs were achieved by step-by-step verification of the front-end design. These designs were successfully tested on the prototyping board of the implemented test platform. This fundamental work is expected to contribute significantly to achieve the single-run VPLC chipset implementation.

References

[1] http://www.smartgrid.or.kr/ [accessed: May 2014]

[2] S. Choi, H. M. Oh, Y. Kim, and Y. H. Kim, “Study on Very High-Rate Power Line Communications for Smart Grid,” KIEE Trans. Inst. Electrical Engineers, vol. 60, no. 6, Jun. 2011, pp. 1255-1260.

[3] S. Eom et al., “Physical layer certification system of power line communication Korea standard,” IEEE International Symposium on Power Line Communications and Its Applications, Apr. 2009, pp. 325-330.

[4] ISO/IEC12139-1, Information technology - Telecommunications and information exchange between systems - Power line communication-High speed PLC medium access control and physical layer-Part1: General requirements, Jul. 2009.

[5] M. Zimmermann and K. Dostert, “A Multipath Model for the Power Line Channel,” IEEE Trans. on Comm., vol. 50, no. 4, Apr. 2002, pp. 553-559.

[6] W. Li, Y. Ma, and L. Wa., “Word Length Estimation for Memory Efficient Pipeline FFT/IFFT Processors,” IEEE Signal Processing Systems, 1999.

[7] T. J. Richardson and R. L. Urbanke, “Efficient Encoding of Low-Density Parity-Check Codes,” IEEE Trans. Inform. Theory, vol. 47, no. 2, Feb. 2001, pp. 638-656.

[8] M. P. C. Fossorier, “Quasi-Cyclic Low-Density Parity-Check Codes From Circulant Permutation Matrices,” IEEE Trans. Inform. Theory, vol. 50, no. 8, Aug. 2004, pp. 1788-1793.

ISBN: 978-1-61208-360-5

Compression of Polysomnographic Signals Using the Discrete Cosine Transform and Deadzone Optimal Quantization

Hugo N. de Oliveira, Arnaldo Gualberto de A. e Silva, Igor C. Diniz, Gustavo B. Sampaio, Leonardo V. Batista

Informatics Department
Federal University of Paraiba (UFPB)
Joao Pessoa, Brazil
Email: {hugoneves, arnaldo.gualberto, ygorcrispim, gustavobrito, leonardo}@ci.ufpb.br

***Abstract*—Data compression techniques for electrocardiographic and electroencephalographic exams have been widely reported in the literature over the last decades; but, there are no papers offering a unique solution for all biological signals typically present in polysomnographic records. Aiming to fill this gap, the present work proposes a method of lossy compression for polysomnographic signals based on optimal quantization of the coefficients obtained from the discrete cosine transform. The potentially grave distortions generated by the information loss are controlled by a compression parameter that may be configured to reach the desired Normalized Percent Root-mean-square Difference generating the optimum quantization vector with a minimization of the Lagrange parameter. The quantized signal is sent to a prediction by partial matching compressor, which works as the entropy coder of this compression strategy. The method was tested using the signals in the Polysomnographic database created by the Massachusetts Institute of Technology and Boston's Beth Israel Hospital, achieving compression ratios between 2.16:1 and 67.48:1 with distortion values between 1.0% and 4.0%.**

Keywords–data compression; telemedicine; polysomnographic signals; lossy compression; discrete cosine transform.

I. INTRODUCTION

The technological advances in data transmission have turned the ability to communicate into one of the foundations of the contemporary society. Access to broadband Internet, despite recent technological advances, still remains as a service which is accessible to few people, especially in third world countries. Likewise, a large amount of hard disk space may represent a great cost for applications with either personal or commercial purposes. One way to ease these problems is to reduce the need for storage and/or transmission of data, while preserving all or most of the information on the original message.

The methods used in messages to minimize the disk space needed for its storage is the process called data compression, and this special type of data processing is classified into lossy and lossless compression techniques. A lossless compression and decompression process of a signal results in a reconstructed signal with exactly the same information as the original one. A lossy compression technique may produce a fairly accurate approximation of the original signal, depending on the compression techniques and the parameters used in these techniques.

The biological signals are among the various types of signals on which lossy compression techniques may be applied. These signals are often used for either biometrical or diagnostic purposes, requiring a very low amount of errors – or even none – in a reconstructed signal decoded after the application of a lossy compression method. Polysomnographic (PSG) monitoring has been useful to clarify the physiological mechanism to produce sleep related signs or attacks, such as apnea, arrhythmia, hemodynamics changes, and/or myocardial ischemia [1]. Therefore, this kind of exam cant be performed during day time and is usually done in specialized clinics. The disk space needed for the storage of one hour of a PSG channel may be as large as 2.57 MB, if a proper digitalization is used in the process. Some PSGs may contain nine data channels, resulting in a disk space of approximately 23 MB/h. In eight hours, the average sleep duration of a human being, this signal requires a space equivalent to 185 MB. This may not seem much for the storage capacity of modern computers, but it represents a huge scalability problem when the exams need to be stored for the rest of the patients life for health progress evaluation. This large amount of space represents a problem for the design of embedded homecare polygraphs, which can perform most of these exams in the patients house. This problem can be eased with the use of smart lossy data compression techniques over the PSGs, achieving a good Compression Ratio (CR).

There are no works describing a solution for a unified method for compression of PSGs; so, some modern techniques for electrocardiographic (ECG) and electroencephalographic (EEG) compression will be described as follows:

The method proposed by Mukhopadhyay et al. [2] uses a differentiation technique to detect all R-peaks in the ECG signal, allowing the algorithm to apply a differential encoding process to R-peak regions (also called QRS regions). The QRS slices are passed to an algorithm based on a Lossless Compression using Character Encoding (LLCCE), since these parts of the ECG are more important to the signal reconstruction. The rest of the ECG is passed to a sub sampler, which reduces the sampling frequency of the signal by one half. Then non-QRS data are processed by a Lossy Compression using Character Coding Encoding (LCCE) scheme. The non-parameterized compressor was tested using the Physikalisch-Technische Bundesanstalt (PTB) Diagnostic ECG Database and reported a CR value of 23.10 : 1 with a Percent Root-mean-square Difference (PRD) value of 7.55%.

Ranjeet et al. [3] uses a Cut And Align (CAB) strategy to slice the ECG signal and reorganize the blocks in a 2D array, which is passed to a 2D Discrete Wavelet Transform (DWT). The remaining coefficients are encoded using Huffman entropy coding, achieving an average of 65% in compression efficiency with 0.999 correlation score. The tests performed using this near lossless strategy resulted in an optimum 2D array size of $180x20$ samples in the Massachusetts Institute of Technology and Boston's Beth Israel Hospital (MIT-BIH) Arrhythmia Database [4].

The work described by Lai et al. [5] explores the use of the Discrete Cosine Transform (DCT) IV – in contrast to the DCT-II, often used for compression purposes. Initially, the ECG signal is divided into DCT blocks with 64 samples, then a differential coding procedure is applied, feeding a Huffman entropy coder. This non-parameterized strategy achieved an average CR of $5.25:1$ with a PRD of 0.19% and a Normalized Percent Root-mean-square Difference (NPRD) of 2.88% using the MIT-BIH Arrhythmia Database [4].

The method described by Anas et al. [6], similarly to the present work, is a DCT-based compressor. It uses the correlation between the ECG cycles (identified by QRS complexes) to eliminate the redundancy in the data of the records. This ECG compressor has a preprocessing step counting with baseline elimination, an average filter, a high pass filter and a butterworth filter, preparing the record to the compression routine. Then, the ECG is passed to a R-peak detector, the ECG cycles are interpolated to a fixed value $M = 512$, normalized to an interval $[0, 1]$, transformed using the DCT, quantized and encoded. This parameterized method achieved good CRs for low PRD values, but only the results obtained by the compression of three records from the MIT-BIH Arrhythmia Database [4] were published.

Srinivasan et al. [7] propose a multichannel near-lossless EEG compression algorithm based on image and volumetric coding. The algorithm arrange multichannel EEG signal in the form 2D image or 3D volume, then apply either a 2D or 3D DWT to exploit simultaneously both spatial and temporal correlation in a single procedure. The proposed algorithm achieved a compression ratio of 6.63 with PRD of 9.21% for a quantizer step-size equals to ten in one of the datasets used.

The fact that no work in the literature describes a compression technique for all the PSGs is responsible for the non-existence of a standard distortion measure for the lossy compression of these signals. However, there is a large amount of papers describing lossy and lossless compression methods for electrocardiographic signals. The lossy compressors often use the PRD as an objective evaluation of the distortion present in the decoded signal. The PRD is defined as:

$$PRD = \sqrt{\frac{\sum_{n=0}^{N-1}\left(x\left[n\right]-\tilde{x}\left[n\right]\right)^2}{\sum_{n=0}^{N-1}\left(x\left[n\right]\right)^2}} \times 100\% \qquad (1)$$

This measure is very sensitive to the baseline of the original signal. A second definition for the PRD, the NPRD, which overcomes this problem, is described by Batista et al. [8] as:

$$NPRD = \sqrt{\frac{\sum_{n=0}^{N-1}\left(x\left[n\right]-\tilde{x}\left[n\right]\right)^2}{\sum_{n=0}^{N-1}\left(x\left[n\right]-\overline{x}\right)^2}} \times 100\% \qquad (2)$$

where $\overline{x}$ is given for:

$$\overline{x} = \frac{1}{N}\sum_{n=0}^{N-1} x\left[n\right] \qquad (3)$$

The NPRD, however, is still not the ideal distortion measure for biological signals, as it does not consider the different characteristics present in each record. This criterion only provides an objective approximation for the amount of errors in the reconstructed signal.

This paper is organized as follows: Section I presents an overview about the PSG data volume problem, the method we propose to solve it and the quality metrics we used. Section II gives an overview about PSG signals. Section III explains the basis of DCT-based lossy compressors. Section IV explains the proposed compression method. Section V shows the results obtained by the test application. Section VI describes the conclusions obtained after the analysis of the results and some possible applications for the proposed method.

II. Polysomnographic Signals

The PSGs may include several types of signals, including both well-known signals as electrocardiograms and electroencephalograms; and signals with more specific purposes, as electrooculograms (EOG), stroke volume (SV) and oxygen saturation records (SO2). The signals included and discarded from the exams are determined by the health condition the physician wants to analyze and the patients health state.

The large diversity of behavior among the PSGs is responsible for the lack of unified compression solutions for all signals described in the biological data compression literature. While some signals are periodical, as ECG, blood pressure (BP) and respiration (Resp) signals, other signals are almost completely chaotic, as electromyographic (EMG), EOG and EEG records. Some works take advantage of the periods in ECG records to achieve greater CR values and other works cover only EEG signals, but none of them was tested in all PSGs.

The work described by Ichimaru and Moody [1] presents a standardized physiological PSG database format, including an amount of 18 signals, with duration ranging from two to seven hours. The PSGs were digitalized with 250 Hz sampling frequency and a 12 bits/sample quantization. The so called MIT-BIH Polysomnographic Database became the standard test corpus for the PSG processing applications. An image of the samples of a signal in the MIT-BIH Polysomnographic Database is shown in Fig. 1.

III. DCT-based Compressors

Among the several domain transformations applicable to digital signals, the DCT [9] and the DWT [10] have been

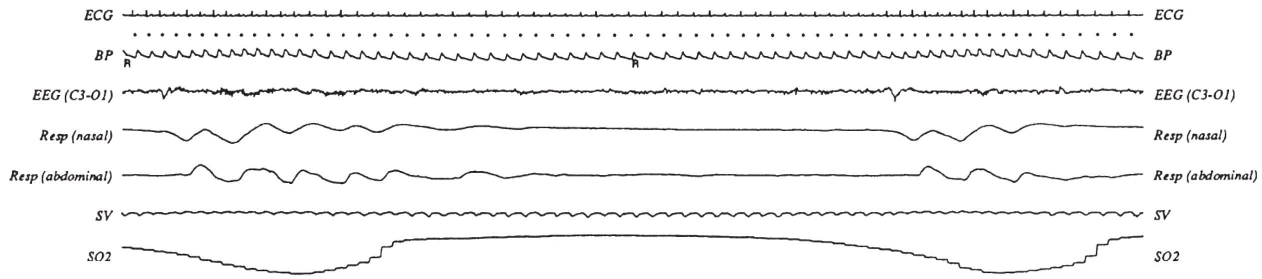


Figure 1. Full disclosure of the polysomography data. Including ECG, BP, EEG, Resp, SV and SO2 records. Adapted from [1].

widely used in lossy compressors due to their energy compaction properties. Fast implementations of both transforms in both 1 and 2 dimensions have been described in the data compression literature over the last decades. The DCT – the most popular one – is used in many encoding formats, including ECG encoders [11][12][13][14]; video encoders [15][16]; still image encoders [17]; and audio encoders [9]. As described by Batista et al. [8], there are four steps often used for the creation of DCT-based encoders to compress a data sequence **x**:

1) Partition of **x** in N_b consecutive blocks $\mathbf{b}_i, i = 0, 1, ..., N_b - 1$, each one with L_b samples;

2) DCT computation for each block;

3) Quantization of the DCT coefficients;

4) Lossless encoding of the quantized DCT coefficients.

Increasing the block size increases the CR and the DCT computing time. Various results show, however, that increasing the block size above a certain point results in a very modest CR gain, while the processing time significantly increases [12][18]. The DCT-II is widely used in lossy data compressors and it is the closest unitary transform approximation for the optimal Karhunen-Love Transform (KLT) [9]. Let $b_i[n], n = 0, 1, ..., L_b - 1$, represent the L_b values in block $\mathbf{b}_i$; the one-dimensional DCT-II of this block generates a transformed block $\mathbf{B}_i$ constituted by a sequence of L_b coefficients $B_i[m], m = 0, 1, ..., L_b - 1$, given by:

$$B_i\left[m\right] = \left(\frac{2}{L_b}\right)^{\frac{1}{2}} c_m \sum_{n=0}^{L_b-1} \left(b_i\left[n\right] \cos\left[\frac{(2n+1)m\pi}{2L_b}\right]\right), \quad m = 0, 1, ..., L_b - 1 \tag{4}$$

where $c_m = 1$ for $1 \leq m \leq L_b - 1$ and $c_0 = \sqrt{\left(\frac{1}{2}\right)}$.

The DCT can be seen as a one-to-one mapping for N-point vectors between the time and the frequency domains [17]. The coefficient $B_i[0]$, which is directly related to the average value of the time-domain block, is often called the DC coefficient, and the remaining coefficients of a block are called AC coefficients.

Given $\mathbf{B}_i$, $\mathbf{b}_i$ can be recovered applying the inverse DCT-II:

$$b_i\left[n\right] = \left(\frac{2}{L_b}\right)^{\frac{1}{2}} \sum_{m=0}^{L_b-1} \left(c_m B_i\left[m\right] \cos\left[\frac{(2n+1)m\pi}{2L_b}\right]\right), \quad n = 0, 1, ..., L_b - 1 \tag{5}$$

To quantize $\mathbf{B}_i$, one can use a quantization vector, q. Each element $q[n], n = 0, 1, ..., L_b - 1$, of q is a positive integer in a specified interval and represents the quantization step size for the coefficient $B_i[n]$. The elements $\hat{B}_i[n]$ of the quantized DCT block $\hat{\mathbf{B}}_i$ are obtained by:

$$\hat{B}_i\left[n\right] = B_i\left[n\right] // q\left[n\right] \tag{6}$$

where the operator // represents the division followed by rounding to the nearest integer.

Ratnakar [19] showed that it is possible to achieve a considerable gain in the CR, for a fixed distortion, by using thresholding. If $t[n], n = 0, 1, ..., L_b - 1$ are the elements of the threshold vector, **t**, the elements of $\hat{\mathbf{B}}_i$ are now given by:

$$\hat{B}_i\left[n\right] = \begin{cases} 0, & \text{if} \left|B_i\left[n\right]\right| < t\left[n\right] \\ B_i\left[n\right] // q\left[n\right], & \text{otherwise} \end{cases}, \quad \begin{array}{rcl} n & = & 0, 1, ..., L_b - 1 \\ i & = & 0, 1, ..., N_b - 1 \end{array} \tag{7}$$

The dequantization, performed during the decompression process to find an approximation to the original coefficients, consists simply in a multiplication of each quantized coefficient by the correspondent component of **q**. For most DCT-based compressors, the quantization is the only lossy operation involved. The definition of **q** and **t** has a strong impact in CR and distortion [19].

Ahmed et al. [13], for example, uses a unique threshold value t_0 for all coefficients. Coefficients with estimated variances less than t_0 are quantized to zero. All elements of the quantization vector are equal to 1. Varying t_0 controls the CR and the distortion.

The CAB/2-D DCT [12] uses a unique quantization step size for all coefficients. This value is defined to minimize the squared mean error between the original and the reconstructed signal, for a given CR. As pointed out by Lee and Buckley [12], the good resulting compression ratios are principally due to a 2D approach, which simultaneously explores the correlation between consecutive samples and consecutive beats of the signal, rather than to the quantization strategy.

The work presented by Poel [11] uses a **q** vector whose components are values from a line segment. The value of $q[0]$ is fixed at 1 and the next values grow linearly up to the value of $q[L_b - 1]$. Varying the inclination of the line segment controls the CR and the distortion.

The lossless encoding of the quantized DCT coefficients generally involves run-length encoding, because the quantization normally generates many null values, followed by an entropy encoder [12].

The present work describes a method to define **q** and **t** in a way that minimizes the estimated entropy of the quantized coefficients for a given distortion, and uses these optimized vectors as the basis for a PSG compressor. The main goal is to demonstrate the possibility of attaining good compression ratios by using a carefully defined quantization strategy.

IV. Description of the Proposed Method

The measure used for the calculation of the distortion after the compression was the NPRD due to the baseline variation among the PSG signals. Some PSGs, such as the respiratory (Resp) signals, may have a very high baseline value, allowing the error amount of the common PRD to grow a lot for a low PRD value. The NPRD unifies the computation of the distortion to a single measure, without the need of a baseline elimination preprocess.

To reduce the long-term storage problem created by exams involving PSGs in small and medium sized clinics, the purposed technique was created. This unified solution was tested in the four main PSG signals: ECG, EEG, BP and Resp. The method works as a parameterized compressor, defining the optimum **q** and **t** vectors for the codification of each channel in the signal. An optimization for the choice of the **q** and **t** parameters was made using the minimization of the Lagrange multiplier for each DCT coefficient, similarly to the work presented in Ratnakar [19], which proposed a solution to the optimum quantization of images. The Lagrangian minimization allows the compressor to perform an independent optimization for each coefficient independently. The number of decoding iterations using exhaustive search methods N_{exa} – as shown in (8) – is then optimized to a much lower value N_{opt}.

$$N_{exa} = {Q_{max}}^B + {T_{max}}^B \tag{8}$$

Lee and Buckley [12] tested the use of a 2D DCT with block sizes from $4x4$ to $64x64$, narrowing the tests only to powers of 2. These tests resulted in a saturation of the coding gain with block sizes around $32x32$ and $64x64$ samples. Based on the experiments presented in [12] and the ones performed by Batista et al. [8], we used block sizes containing 16, 32 and 64 samples in the tests.

For a given signal, let $H(\mathbf{q},\mathbf{t})$ be the zero-order entropy of all DCT coefficients quantized by using **t** and **t**, and $D(\mathbf{q},\mathbf{t})$ a measure of the distortion introduced in the PSG signal by the quantization. The proposed optimization problem can then be given by the statement: for a given $D(\mathbf{q},\mathbf{t})$, determine **q** and **t** in a way that minimizes $H(\mathbf{q},\mathbf{t})$.

Optimization can be achieved by minimizing the Lagrangian $J = H(\mathbf{q},\mathbf{t}) + \lambda D(\mathbf{q},\mathbf{t})$ for a given value of the Lagrange multiplier λ [19]. The value of λ that leads to the desired $H(\mathbf{q},\mathbf{t})$ or $D(\mathbf{q},\mathbf{t})$, within a given tolerance, can be efficiently found by using the bisection method [20]. The Lagrangian minimization allows the compressor to set a maximum number of decoding operations empirically predefined by tests of the method on the signals. The number N_{opt} was set to the value 17 in the test application, but this number may vary according to the type of signal, its digitalization parameters and its statistic distribution.

Empirical tests showed that NPRD values lower than or equal to 3.0% required Q_{max} and T_{max} values lower than 128, thus, this was the maximum value the elements of the QT vectors could reach for these distortions. For NPRD values higher than 3.0%, the maximum values for the QT elements were set to 256. Therefore, for NPRD values higher than 3.0% using 64 samples DCT blocks, N_{exa} would assume the value $2.68x10^{154}$. The process described in the next paragraphs allows reducing the complexity of the problem to practical levels. It should be noted that the entire records to be compressed are used to calculate the optimal **q** and **t** vectors.

For the optimization procedure, we use the mean square error as the distortion measure $D(\mathbf{q},\mathbf{t})$. Since the DCT is an orthonormal transform, $D(\mathbf{q},\mathbf{t})$ can be calculated from the distortions introduced in the DCT coefficients [19]. This eliminates the need to apply the inverse DCT to the dequantized coefficients in order to measure the distortion in the time-domain. Thus, the mean squared error introduced by the quantization can be calculated as:

$$D\left(\mathbf{q},\mathbf{t}\right) = \frac{\sum_{i=0}^{N_b-1}\left[\sum_{n=0}^{L_b-1}\left(B_i\left[n\right] - q\left[n\right]\hat{B}_i\left[n\right]\right)^2\right]}{L_b N_b} \tag{9}$$

The mean square error due to the quantization of coefficient number k of all blocks, which will be called $D_k(q[k],t[k])$, is given by:

$$D_k\left(q\left[k\right],t\left[k\right]\right) = \frac{1}{N_b}\sum_{i=0}^{N_b-1}\left(B_i\left[k\right] - q\left[k\right]\hat{B}_i\left[k\right]\right)^2 \tag{10}$$

Thus, we can write (9) as:

$$D\left(\mathbf{q},\mathbf{t}\right) = \frac{\sum_{n=0}^{L_b-1} D_n\left(q\left[n\right],t\left[n\right]\right)^2}{L_b N_b} \tag{11}$$

Consider now that the coefficient number k of the quantized blocks assumes value v in $n_k(v)$ of the N_b blocks. Then the entropy $H_k(q[k],t[k])$ of the coefficient number k measured over all quantized DCT blocks is given by:

$$D_k\left(q\left[k\right],t\left[k\right]\right) = -\sum_v\left[p_k\left(v\right)\log_2\left(p_k\left(v\right)\right)\right] \tag{12}$$

where $p_k(v) = n_k(v)/N_b$.

To estimate the entropy of all quantized coefficients we use the following simplified model [19]:

$$H\left(\mathbf{q},\mathbf{t}\right) = \frac{1}{L_b}\sum_{n=0}^{L-1}\left[H_n\left(q\left[n\right],t\left[n\right]\right)\right] \tag{13}$$

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-360-5

In the experimental results presented by Ratnakar [19], the error between the estimated and the real entropy was normally below 0.02 bits/symbol, which indicates the precision of the model.

With the possibility to calculate $D(\mathbf{q},\mathbf{t})$ and $H(\mathbf{q},\mathbf{t})$ as the mean of the distortion and of the entropy of each coefficient, the minimization of J reduces to the minimization of:

$$\begin{gathered} J_n = H_n\left(q\left[n\right], t\left[n\right]\right) + \lambda D_n\left(q\left[n\right], t\left[n\right]\right), \\ n = 0, 1, ..., L_b - 1 \end{gathered} \tag{14}$$

In other words, the minimization can be done independently for each coefficient. With this simplification, if $L_b = 64$ samples and the elements of $\mathbf{q}$ are integer values in the range 1 to 256, only $64 * 256 = 2^{14}$ of the 64^{256} possible values of $\mathbf{q}$ need to be analyzed in the minimization procedure. This complexity reduction combined with the use of histograms, incremental calculations and other techniques [19], allow performing an efficient search for the optimum $\mathbf{q}$ and $\mathbf{t}$ vectors.

After defining the optimum $\mathbf{q}$ and $\mathbf{t}$ for a given signal, the compressor closely follows the steps of general DCT-based compressors already described. A scheme representing graphically the compressors steps is shown in Fig. 2.

The dead-zone quantization step of the encoding process normally generates a large amount of subsequent null frequency samples, mostly in the high frequency AC coefficients. This characteristic of the quantized signal allows the compressor use an efficient lossless entropy coding technique, such as a Golomb coding [21] or a Huffman coding [22]. The proposed method does not aim to achieve a hardware implementation compression model; so, for validation purposes, we used a more efficient lossless coding algorithm, the Prediction by Partial Matching (PPM) [23]. The optimal entropy coding of the DC coefficients – which tend to assume higher values than the AC coefficients – was also decisive to the choosing of the PPM. The PPM creates a generic statistic distribution model in the coding process, so the high values in the DC coefficients are not a problem for the optimal coding of the signal.

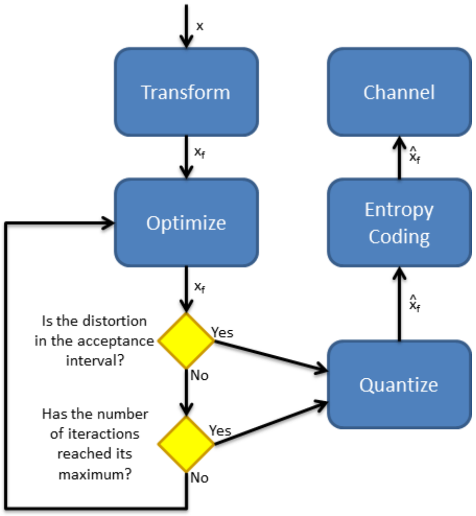

x
Transform
x_f
Optimize
x_f
Is the distortion in the acceptance interval?
Yes
No
Has the number of iteractions reached its maximum?
Yes
No
Quantize
$\hat{x}_f$
Entropy Coding
$\hat{x}_f$
Channel

Figure 2. Scheme showing the processing steps used by the proposed compressor.

The decompressor, as in many other lossy codecs, is a simple subset of the compressor. It is composed by three simple steps, recreating an approximation to the original signal by applying the inverse encoding operations in an inverse order. The first decompression stage is to run a decoding operation, retrieving the domain frequency quantized signal $\hat{\mathbf{x}}_f$ from the channel it was stored by the compressor. Then a dequantization operation is applied, creating the approximation to the signal still composed by frequency domain coefficients $\tilde{\mathbf{x}}_f$. At last, an inverse DCT transform is run over the DCT blocks in the $\tilde{\mathbf{x}}_f$ signal, generating the approximated time domain original signal $\tilde{\mathbf{x}}$. Fig. 3 presents an overview of the decompression scheme.

V. Results and Discussion

Table I and Table II show, respectively, the CR results from compression– using NPRD values among 1.0% and 4.0%, and with DCT block sizes of 16, 32 and 64 samples – of ECG and EEG signals, which showed the best visual reconstruction qualities. Table III and Table IV show the findings for BP and Resp signals, respectively, which obtained good results with certain combinations of parameters, although they have tolerated NPRD values lower than the other signals to approximately the same level of visual distortion. The thresholding effect of these signals proved to be stronger than the other PSGs, explaining the high CR values obtained, since most of the information to be encoded by the PPM in the entropy coding stage is concentrated on the DC levels of the blocks generated by the DCT signals.

In more chaotic signals, which also have a larger amount of noise, such as ECGs, the best results of optimum compression were obtained with DCT blocks of smaller size, as seen in Fig. 4(a). This result is reversed in the case of more linear and less noisy PSGs, such as BP and Resp signals, which is shown in Fig. 4(b).

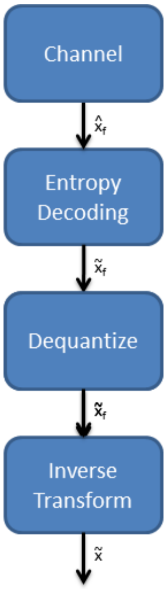


Figure 3. Scheme showing the processing steps used by the proposed decompressor.

TABLE I. CR RESULTS FOR NPRD VALUES BETWEEN 1.0% AND 4.0% AND DCT BLOCK SIZES OF 16, 32 AND 64 SAMPLES FOR ECG SIGNALS.

DCT block size	NPRD						
	1.0%	**1.5%**	**2.0%**	**2.5%**	**3.0%**	**3.5%**	**4.0%**
16	2.52401	3.05967	3.49161	4.0208	4.8165	5.22089	5.98384
32	2.43439	2.8334	3.23771	3.84271	4.38118	4.83237	5.22027
64	2.26796	2.64573	3.02288	3.43338	3.84632	4.09859	4.72272

TABLE II. CR RESULTS FOR NPRD VALUES BETWEEN 1.0% AND 4.0% AND DCT BLOCK SIZES OF 16, 32 AND 64 SAMPLES FOR EEG SIGNALS.

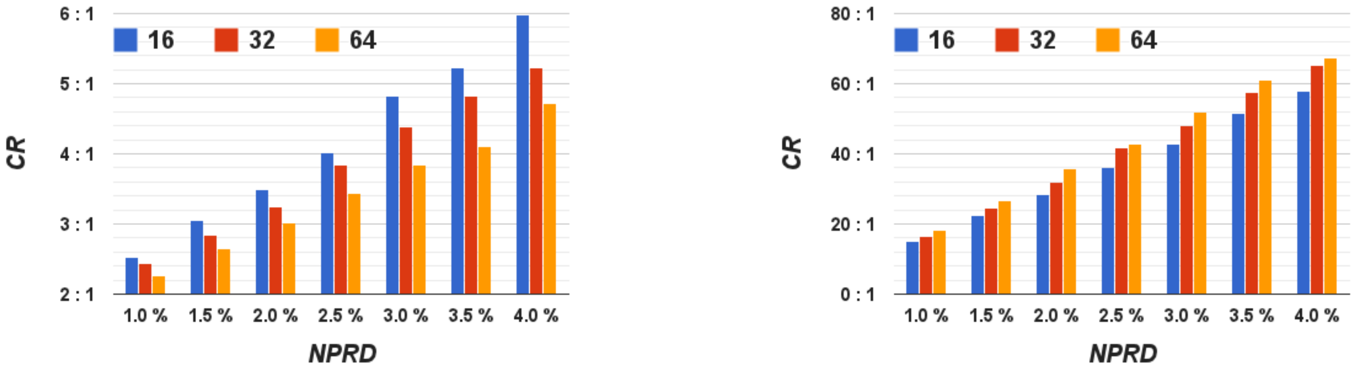
DCT block size	NPRD						
	1.0%	**1.5%**	**2.0%**	**2.5%**	**3.0%**	**3.5%**	**4.0%**
16	2.16316	2.49929	2.78552	3.05686	3.34842	3.6387	3.9742
32	2.18337	2.53933	2.82184	3.13717	3.43413	3.79547	4.18758
64	2.21229	2.55054	2.8828	3.09125	3.53912	3.75921	4.07101

TABLE III. CR RESULTS FOR NPRD VALUES BETWEEN 1.0% AND 4.0% AND DCT BLOCK SIZES OF 16, 32 AND 64 SAMPLES FOR BP SIGNALS.

DCT block size	NPRD						
	1.0%	**1.5%**	**2.0%**	**2.5%**	**3.0%**	**3.5%**	**4.0%**
16	7.05813	9.83612	12.5254	15.0561	16.9096	18.5575	20.813
32	7.6318	11.0152	13.7244	16.2092	18.397	20.6467	22.6501
64	8.31705	11.9357	14.9918	17.6552	20.0764	22.2745	24.4644

TABLE IV. CR RESULTS FOR NPRD VALUES BETWEEN 1.0% AND 4.0% AND DCT BLOCK SIZES OF 16, 32 AND 64 SAMPLES FOR RESP SIGNALS.

DCT block size	NPRD						
	1.0%	**1.5%**	**2.0%**	**2.5%**	**3.0%**	**3.5%**	**4.0%**
16	14.9775	22.3109	28.533	36.1156	42.9171	51.4388	57.9435
32	16.4825	24.6164	32.0925	41.7645	47.9586	57.3724	65.4241
64	18.2037	26.8057	35.7891	42.893	52.0199	61.0977	67.4799



(a)

(b)

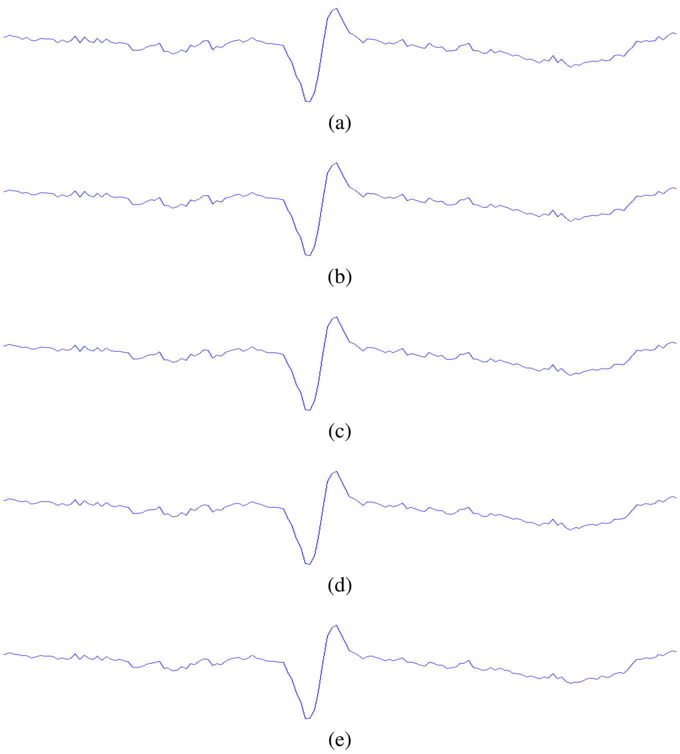
Figure 4. NPRD x CR graphic showing the evolution of the CRs for different block sizes in (a) ECG signals. (b) Resp signals.

As seen in Fig. 5, the reconstruction of the ECG signals achieved very good results, even for higher NPRD values, as 4.0%. This means that this value can be increased even more without grave reconstruction errors. EEG signals, although much more chaotic, had a reconstruction quality close to the ECGs for the same values of NPRD, as seen in Fig. 6. As expected, these signals obtained the lowest CRs among the PSGs due to the large amount of information present in their samples.

The reconstruction of BP signals, exemplified by Fig. 7, showed a strong thresholding effect for signal reconstruction with NPRD values higher than 3.0%. The compression of these signals resulted in high CRs, even for tests with lower NPRD values, allowing compressions with milder distortions to be applied and still result in acceptable CR values. Some kinds of medical exams require only the basic shape of the signal to be stored, so, depending on the purpose of the exam, the thresholding effect can be accepted for BP signals, achieving higher CRs.

In Resp signals (see Fig. 8) like the BPs, the effect of thresholding was strong enough in tests with higher NPRD values, but the CR obtained with the compression of these PSGs is the best among all PSGs thanks to its continuity and to a small amount of noise present in the samples. This allows the application of lighter quantization, still achieving good CR values.

Compression using the Lagrangian minimization to determine the optimal parameters did not behave well with signals



(a)

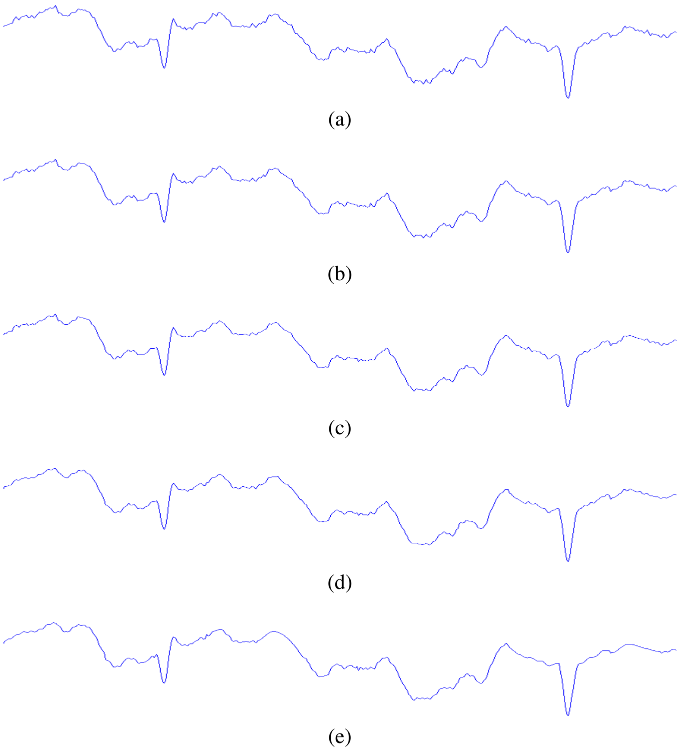
(b)

(c)

(d)

(e)

Figure 5. Slice of the ECG channel of the file slp37.dat of the MIT-BIH Polysomnographic Database (a) original. (b) reconstructed with a NPRD of 1.0%. (c) reconstructed with a NPRD of 2.0%. (d) reconstructed with a NPRD of 3.0%. (e) reconstructed with a NPRD of 4.0%.



(a)

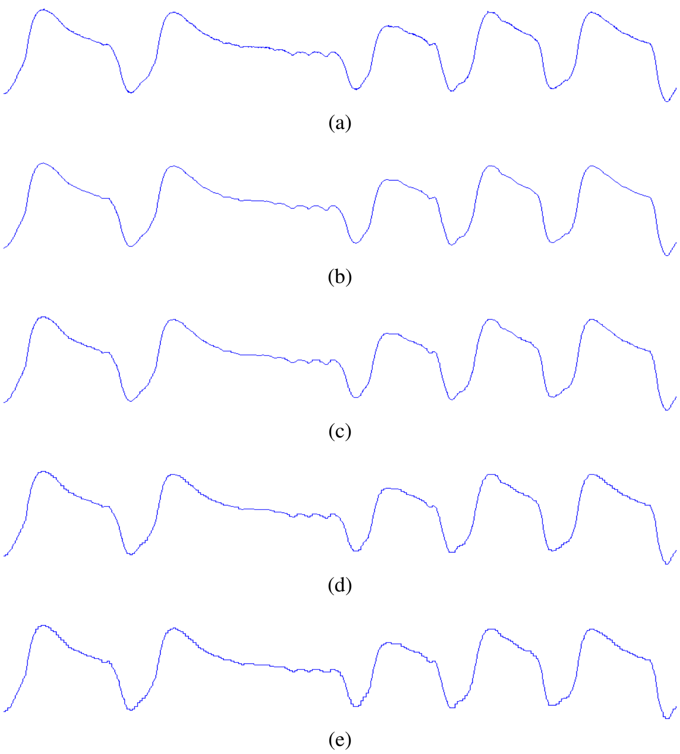
(b)

(c)

(d)

(e)

Figure 6. Slice of the EEG channel of the file slp01a.dat of the MIT-BIH Polysomnographic Database (a) original. (b) reconstructed with a NPRD of 1.0%. (c) reconstructed with a NPRD of 2.0%. (d) reconstructed with a NPRD of 3.0%. (e) reconstructed with a NPRD of 4.0%.



(a)

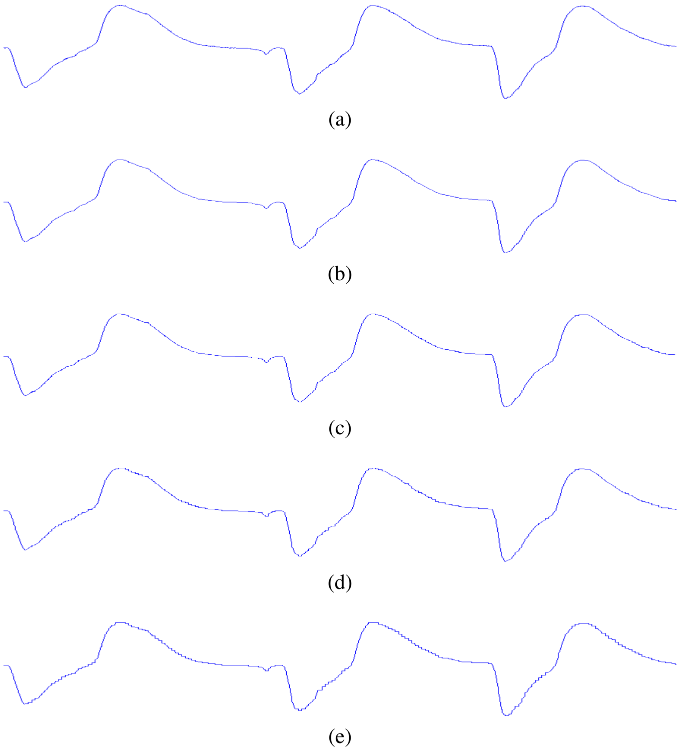
(b)

(c)

(d)

(e)

Figure 7. Slice of the BP channel of the file slp16.dat of the MIT-BIH Polysomnographic Database (a) original. (b) reconstructed with a NPRD of 1.0%. (c) reconstructed with a NPRD of 2.0%. (d) reconstructed with a NPRD of 3.0%. (e) reconstructed with a NPRD of 4.0%.



(a)

(b)

(c)

(d)

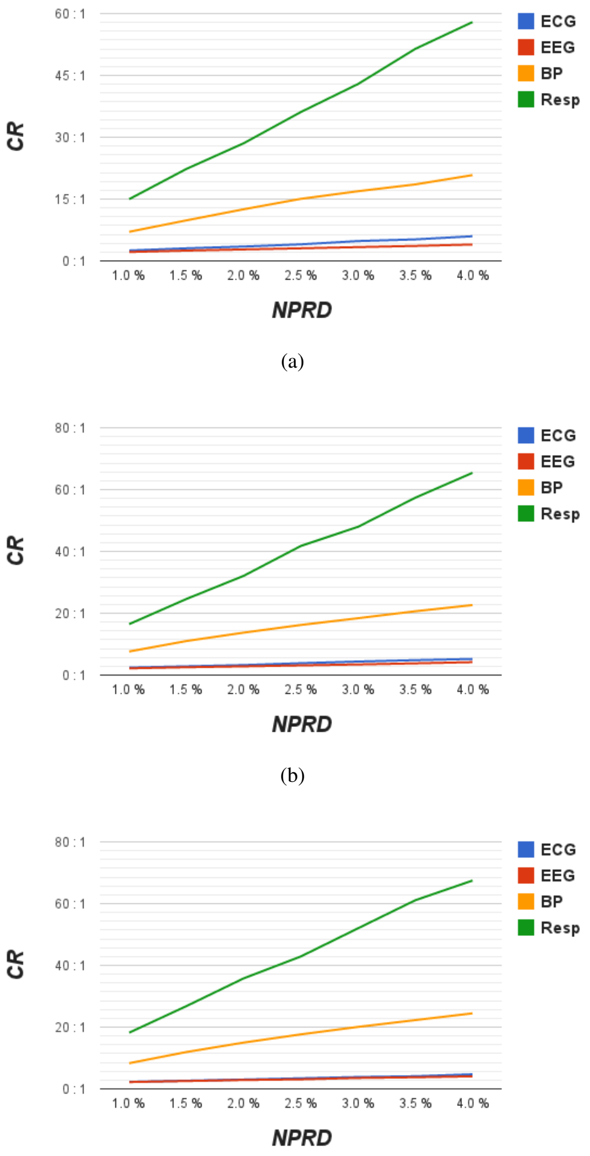
(e)

Figure 8. Slice of the Resp channel of the file slp16.dat of the MIT-BIH Polysomnographic Database (a) original. (b) reconstructed with a NPRD of 1.0%. (c) reconstructed with a NPRD of 2.0%. (d) reconstructed with a NPRD of 3.0%. (e) reconstructed with a NPRD of 4.0%.

ISBN: 978-1-61208-360-5

with a highly uneven statistical distribution. The large number of samples with values outside the normal range interferes with the calculation of the NPRD, since this measure is inversely proportional to the standard deviation, which is highly affected by values outside a certain range near the baseline of the signal. A high standard deviation allows a higher value for the numerator of the equation – the Root Mean Square Error (RMSE) – to achieve the same NPRD, which implies a very large amount of visual errors.

Although in most signals the DCT blocks with 64 samples obtained the best CR, these blocks also feature a worse visual reconstruction quality, if compared to PSGs with the same NPRD values compressed using blocks with 16 or 32 samples. The CR evolution graphics for different NPRD values in ECG, EEG, BP and Resp records are seen in Fig. 9.



(a)

(b)

(c)

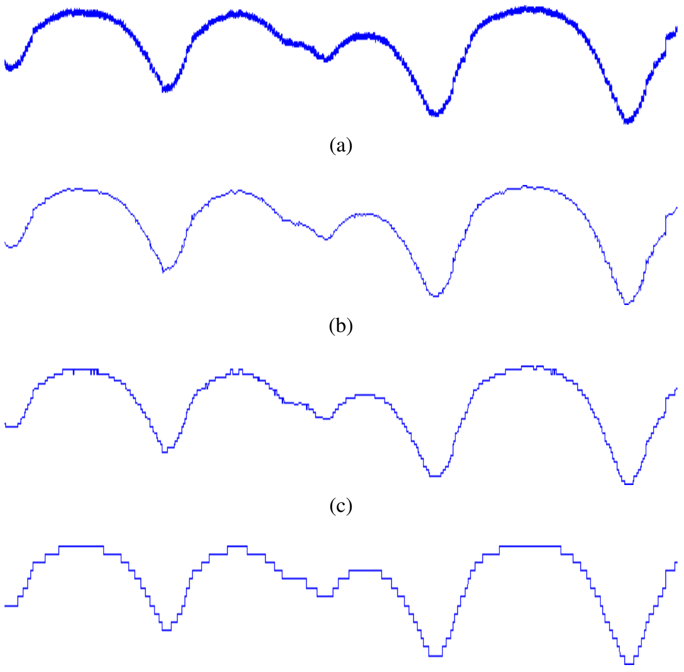
Figure 9. CR evolution for ECG, EEG, BP and Resp signals according to different NPRD values in DCT blocks of (a) 16 samples. (b) 32 samples. (c) 64 samples.

All PSGs showed basically the same CR evolution in different blocks sizes. In some signals it is noticeable a greater change in the CR values for higher NPRDs, what may be attributed to the amount of redundant information, the amount of noise present in the original signal and the amplitude of the samples.

The optimal quantization tends to eliminate signal noise, leaving its baseline more visible and removing the temporal redundancy present in their samples. Some signals – as Resp and BPs – suffer from a faster saturation in the visual quality because of their higher standard deviation.

Depending on the amount of noise that can be accepted in the ECG and EEG signals, an increase of desired NPRD passed to the compressor presents a good CR performance at the cost of low noise, reaching $5.98 : 1$ and $4.18 : 1$, respectively, for NPRD values of 4.0%. The CR values of ECGs came close to the compression obtained by [8], if a visual analysis of the reconstructed signals is performed. A more detailed comparison with studies involving the compression of ECGs and EEGs is hampered by the use of databases where the digitalization process was done differently than [1], hindering the comparison using objective distortion measures. The frequent use of the PRD – and not the NPRD – for the calculation of the errors of the reconstructed signal is also a factor that complicates a comparison with other papers.

Fig. 10 shows the effect of sensor defects in a SO2 signal. These PSG channels – along with the EMGs, EOGs and SV – were not considered by the tests because of the large amount of errors present in their capture processes and the low number of signals containing channels of these records in [1].



(a)

(b)

(c)

(d)

Figure 10. Slice of the SO2 channel of the file slp67x.dat of the MIT-BIH Polysomnographic Database affected by errors (a) original. (b) reconstructed with a NPRD of 1.0%. (c) reconstructed with a NPRD of 2.0%. (d) reconstructed with a NPRD of 4.0%.

The runtime of the compressor varied according to the length of the signal, since the database has no fixed length for the signals. The larger signals considered for this papers test routines – with six hours and four data channels – were compressed in less than 30 minutes. This result enables the adoption of the proposed method for practical cases, since the compression method took less than a sixth of the length of the signals to compress the data. Since most polysomnographic exams are performed during night time, clinics can use a fraction of the daytime to compress the PSG records captured during the previous night.

Embedded mobile systems for PSG exams with less processing power than conventional computers may divide the captured recordings into slices, allowing the execution of the record to run at the same time as the compress algorithm for the previous data section. This may result in cheaper homecare PSG capture devices, bringing more comfort to the patients.

VI. Conclusion

Technological advances in the processing and storage capacities of personal computers and the price reduction of the biological signal sensors allowed the popularization of medical exams using polysomnographic signals. There was an increase in the data volume generated by these types of exams, although no compression techniques covering the codification of all PSGs have been reported in the data compression literature.

We presented a lossy parameterized compression method for PSGs, prioritizing the reconstruction quality of the compressed signals. A Lagrangian minimization was used to drastically reduce the computational complexity for the choice of the optimal dead-zone quantization vectors. The amount of errors allowed in the reconstruction of the PSGs may vary according to their diagnostic purpose, presetting the compressor in order to tolerate lower or higher objective distortions.

For low NPRD values the compressor achieved different results, depending on the entropy of the PSG. The best CRs were reached in EEG signals, which varied between $2.16 : 1$ and $4.17 : 1$. In ECG signals, on the other hand, CR ranged between $2.26 : 1$ and $5.98 : 1$. The BP signals were compressed with CRs in interval of $7.05 : 1$ and $24.06 : 1$. The highest compression ratios were obtained by the method in Resp signals, with values in the range of $14.97 : 1$ and $67.47 : 1$. However, low objectives distortion metrics do not imply in a good visual quality of signals reconstruction.

Empirical tests validated the presented codification technique, which achieved a small runtime in comparison to the length of the original signals. The simplicity of the method may be a motivation for the development of both hardware implemented solutions and desktop applications.

References

[1] Y. Ichimaru and G. B. Moody, "Development of the polysomnographic database on cd-rom," Psychiatry and Clinical Neurosciences, vol. 53, no. 2, 1999, pp. 175–7. [Online]. Available: http://www.biomedsearch.com/nih/Development-polysomnographic-database-CD-ROM/10459681.html

[2] S. Mukhopadhyay, M. Mitra, and S. Mitra, "An ecg data compression method via rpeak detection and ascii character encoding," in Computer, Communication and Electrical Technology (ICCCET), 2011 International Conference on, March 2011, pp. 136–141.

[3] K. Ranjeet, A. Kumar, and R. Pandey, "An efficient compression system for ecg signal using qrs periods and cab technique based on 2d dwt and huffman coding," in Control, Automation, Robotics and Embedded Systems (CARE), 2013 International Conference on, Dec 2013, pp. 1–6.

[4] G. Moody and R. Mark, "The impact of the mit-bih arrhythmia database," Engineering in Medicine and Biology Magazine, IEEE, vol. 20, no. 3, May 2001, pp. 45–50.

[5] S.-C. Lai, C.-S. Lan, and S.-F. Lei, "An efficient method of ecg signal compression by using a dct-iv spectrum," in Communications, Circuits and Systems (ICCCAS), 2013 International Conference on, vol. 1, Nov 2013, pp. 46–49.

[6] E. Anas, M. Hossain, M. Afran, and S. Sayed, "Compression of ecg signals exploiting correlation between ecg cycles," in Electrical and Computer Engineering (ICECE), 2010 International Conference on, Dec 2010, pp. 622–625.

[7] K. Srinivasan, J. Dauwels, and M. R. Reddy, "Multichannel eeg compression: Wavelet-based image and volumetric coding approach," IEEE Journal of Biomedical and Health Informatics, 2013, pp. 113–120.

[8] L. V. Batista, E. U. K. Melcher, and L. C. Carvalho, "Compression of ecg signals by optimized quantization of discrete cosine transform coefficients," Medical Engineering & Physics, vol. 23, no. 2, 2001, pp. 127 – 134. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1350453301000303

[9] K. R. Rao and P. Yip, Discrete Cosine Transform: Algorithms, Advantages, Applications. San Diego, CA, USA: Academic Press Professional, Inc., 1990.

[10] G. Strang and T. Nguyen, Wavelets and Filter Banks. Wellesley-Cambridge Press, 1996. [Online]. Available: http://books.google.com.br/books?id=Z76N_Ab5pp8C

[11] J. v.d. Poel, "Eletrocardiogram signals compression," Master's thesis, NETEB/UFPB, João Pessoa, Brazil, May 1999.

[12] H. Lee and K. M. Buckley, "Ecg data compression using cut and align beats approach and 2-d transforms," IEEE Transactions on Biomedical Engineering, vol. 46, no. 5, 1999, pp. 556–64. [Online]. Available: http://www.biomedsearch.com/nih/ECG-data-compression-using-cut/10230134.html

[13] N. Ahmed, P. J. Milne, and S. G. Harris, "Electrocardiographic data compression via orthogonal transforms," Biomedical Engineering, IEEE Transactions on, vol. BME-22, no. 6, Nov 1975, pp. 484–487.

[14] F. Zou and R. R. Gallagher, "Ecg data compression with wavelet and discrete cosine transforms," Biomedical Sciences Instrumentation, vol. 30, 1994, pp. 57–62. [Online]. Available: http://www.biomedsearch.com/nih/ECG-data-compression-with-wavelet/7948650.html

[15] Recommendation H.264, ISO/IEC and ITU-T, 2003.

[16] Recommendation H.262, ISO/IEC and ITU-T, 1995.

[17] G. K. Wallace, "The jpeg still picture compression standard," Communications of the ACM, 1991, pp. 30–44.

[18] M. Nelson and J.-L. Gailly, The Data Compression Book (2nd Ed.). New York, NY, USA: MIS:Press, 1996.

[19] V. Ratnakar, "Quality-controlled lossy image compression," Tech. Rep., 1997.

[20] A. Ortega, Optimization Techniques for Adaptive Quantization of Image and Video Under Delay Constraints. Columbia University, 1994. [Online]. Available: http://books.google.com.br/books?id=PQR8HwAACAAJ

[21] S. Golomb, "Run-length encodings," Information Theory, IEEE Transactions on, vol. 12, no. 3, 1966, pp. 399–401.

[22] D. A. Huffman, "A method for the construction of minimum-redundancy codes," Proceedings of the Institute of Radio Engineers, vol. 40, no. 9, September 1952, pp. 1098–1101. [Online]. Available: http://compression.graphicon.ru/download/articles/huff/huffman_1952_minimum-redundancy-codes.pdf

[23] J. G. Cleary and I. H. Witten, "Data compression using adaptive coding and partial string matching," IEEE Transactions on Communications, vol. 32, no. 4, April 1984, pp. 396–402.

Standards for Cooperative Intelligent Transportation Systems: a Proof of Concept

Rodrigo Silva, Satoru Noguchi,
Thierry Ernst, Arnaud de La Fortelle
Mines ParisTech
France
rodrigo_silvabr@yahoo.com, satoru.noguchi@mines-paristech.fr
{thierry.ernst, arnaud.de_la_fortelle}@mines-paristech.fr

Walter Godoy Junior
Federal University of Technology Paraná - UTFPR
Brazil
godoy@utfpr.edu.br

***Abstract*—In recent years, a wide variety of stakeholders have been working for the development of Intelligent Transportation System solutions. Cooperation among the various actors of transportation (vehicles, but also pedestrians, roads and infrastructure, traffic control centers, etc.) is seen as promising to enhance the efficiency of transportation and reduce its negative impacts (e.g., fatalities). However, it means that all communicating entities have to talk the same language, hence the need for Cooperative Intelligent Transportation Systems standards. There are now lots of standards being produced by standardization organization, e.g., International Standardization Organization (ISO) and European Telecommunications Standards Institute (ETSI) and there is a real need to understand how these standards can be implemented. This paper overviews the Intelligent Transportation System station reference architecture and presents a way of practical implementation of a toy Android application based on these standards as a proof of concept implementation. To our knowledge, this is the first implementation description compliant with these standards.**

***Keywords*—*Cooperative Intelligent Transportation System; ITS; Standards; ISO; Wireless Networks*.**

I. Introduction

Transportation systems are increasingly stressed all around the globe, especially in urban areas, and there is a clear need to optimize them. An Intelligent Transportation System (ITS) is seen as a solution to provide innovative services relating to different modes of transport and traffic management. The intelligence is brought by the ability of the system to react using sensors and information processing. Most of the ITS systems deployed today are autonomous in the sense that they are stand-alone and dedicated systems (e.g., traffic lights at an intersection, smart braking systems in a vehicle, etc.). According to the functionality and their purpose, the ITS applications can be classified in three primary categories [1]:

- **Safety:** Improve driving safety, e.g., preventing collision and accident reporting;
- **Efficiency:** Traffic monitoring and traffic management;
- **Infotainment:** Video streaming and Internet access.

Tee next step is to connect these systems through communication and having them cooperate, at least the two first mentioned above (safety and efficiency) [2] [3]. In recent years, a wide variety of stakeholders have been working for development of ITS communication, such as the CAR-2-CAR Communication Consortium gathering most of the European car makers and suppliers, the European Commission through several research projects ([4] [5] [6]), US DoT and Japan. Aside the need for efficient communication, despite limited bandwidth provided by physical carrier, one of the most important things is the interoperability, because system components can be developed by different stakeholders and ITS system shall support modular-based integration.

There are mainly two ways to ensure interoperability: industry standards or open standards produced by standardization organization such as ISO, ETSI, Internet Engineering Task Force (IETF), Institute of Electrical and Electronics Engineers (IEEE) or European Committee for Standardization (CEN). Our work is based on the second option and refers mainly to the common ITS architecture designed by ISO 21217 [7]. This architecture is the basis for several standards within ISO and beyond (ETSI and CEN notably) and a set — hopefully consistent — of ITS standards is under developments.

However, standards never look like a developers guide and they give some room for freedom in the way to implement. To the opposite, all standards are not produced by the same people and always result from compromises, so that there is no guarantee all standards are consistent even though great efforts are made to do so. Therefore, it is of high importance to understand what are the relevant standards for a given application, how they can be translated into functional components and what are the choices a designer of a cooperative application can safely do.

Based on a set of the ITS standards mentioned on Section II, this paper presents a way of practical implementation of the necessary ITS functions, why these set was chosen. We intentionally choose a very simple application (position sharing application) since the focus is not on the application part but on the underlying functions described by the standards: we refer to the implementation work that is "below" the application. To demonstrate this implementation, an Android application was created allowing several mobiles to exchange information (i.e., the location of each mobile). Within each mobile, an interface represents one's own location and the nearby mobiles' location. To our knowledge, this is the first

ISBN: 978-1-61208-360-5

implementation based on the ITS Standards and it shows a proof of concept, demonstrating how ITS standards can actually be implemented.

The paper is organized as follows. Section II overviews the related ITS standards that have been used for this work. The potential system architecture is described in Section III; then, Section IV details the implementation of the ITS standard-compliant application on Android. After discussing the outcomes and potential issues of this development in Section V, Section VI concludes the paper and proposes future directions.

II. Relevant standards

The ISO 21217 standard *ITS - Communications access for land mobiles (CALM) - Architecture* [7] is fundamental for our application since it gives the reference frame for our implementation and the other ITS standards refer to it. It was prepared by Technical Committee ISO/TC 204, ITS subcommittee and describes the communications reference architecture of nodes called *ITS stations* designed for deployment in ITS communication networks.

Figure 1 shows the general ITS Station reference architecture, including interfaces between the various blocks with informative details.

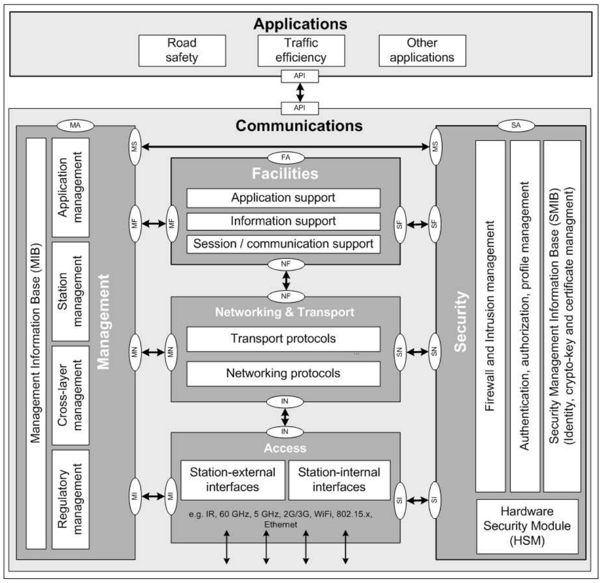

Applications
Road safety
Traffic efficiency
Other applications
API
API
Communications
MA
SA
MS
MS
FA
Facilities
Application support
Information support
Session / communication support
MF
MF
SF
SF
NF
NF
Networking & Transport
Transport protocols
Networking protocols
MN
MN
SN
SN
IN
IN
Access
Station-external interfaces
Station-internal interfaces
e.g. IR, 60 GHz, 5 GHz, 2G/3G, WiFi, 802.15.x, Ethernet
MI
MI
SI
SI
Management Information Base (MIB)
Application management
Station management
Cross-layer management
Regulatory management
Management
Security
Firewall and Intrusion management
Authentication, authorization, profile management
Security Management Information Base (SMIB) (Identity, crypto-key and certificate managment)
Hardware Security Module (HSM)

Fig. 1. ITS Station reference architecture [7].

The ITS architecture [7] is composed by: **"Access"** layer, comprised of OSI layers 1 (Physical) and 2 (Data Link); **"Networking & Transport"** layer, comprised of OSI layers 3 (Network) and 4 (Transport); **"Facilities"** layer, provides application, information and communication supports and it is comprised of OSI layers 5 (Session), 6 (Presentation) and 7 (Application); **"Management"**, a cross entity that containing station management functionality; **"Security"**, a cross entity that provide security services to others entities and **"Application"**, a horizontal entity, which provides Human-Machine Interface.

The ISO/TS 17423 standard *ITS Cooperative systems - ITS application requirements for automatic selection of communication interfaces* [8] is relevant for our application because it specifies the requirements which we will use (e.g, FlowType and transmission/reception Port Number). It relates to ISO 21217 describing the ITS application requirements for automatic selection of communication interfaces by *System Management* entity. To select this communication profile, the *System Management* entity uses the communication requirements, objectives of applications, communication protocol status, regulations and policies. The requirements are divided on five main classes: Operational, Destination, Performance, Security, and Protocol.

The ISO/TS 17419 standard *ITS Cooperative systems - Classification and management of ITS applications in a global context* [9] is used for our application to give the identifiers for each application process or entity. It illustrates and specifies global classification and management of ITS applications.

The ISO/NP 17429 *ITS - Cooperative systems - Profiles for processing and transfer of information between ITS stations for applications related to transport infrastructure management, control and guidance* [10] is necessary for our application since it gives us the procedure to exchange data between our mobile devices. It defines procedures useful to designers and developers of ITS applications exchanging data between ITS stations based on the ITS station reference architecture (ISO 21217).

The ISO 24102-3 standard *ITS Communications access for land mobiles (CALM) - ITS station management Part 3: Service access points* [11] is used for our application to implement the service access points. It specifies the management service access points between the entities and layers described by ISO 21217 (e.g., Management, Facilities, Access, etc.).

III. Design of a Test ITS Application

To evaluate how ITS standards can be implemented, we develop a toy ITS application on personal mobile devices in compliance with the ITS station reference architecture. We implement a simple *position sharing* application, tested with pedestrians, which sends its position and shows neighboring pedestrians' location on Android devices.

In this section, at first, we show the basic requirements of our application, then describe a number of design choices to adapt the application to ITS standards. Because of a wide variety of ITS standards, there are two essential decisions: (i) *which standards should be implemented to satisfy application's requirement*, and (ii) *how they can be adapted to actual implementation*. This section, therefore, explores a set of key ITS standards and design choices to implement them.

A. Application requirements

The proposed application is designed to detect nearby pedestrians with smartphones and/or tablets, and to detect and show the geographic position of surrounding devices.

ISBN: 978-1-61208-360-5

To realize it, firstly, the application needs to obtain device's current location, e.g., from GPS. At the same time, it is necessary to detect nearby devices and exchange positions with each other. The received position information shall be time-stamped and validated, then displayed on a graphical interface. Regarding the communication, from the application's point of view, any type of access technology and protocol may be used as long as the location information can be exchanged.

In summary, the functional requirements of the proposed application are as follows:

- **Location management**: obtain device's own location
- **Discovery**: detect the presence of nearby devices
- **Communication**: exchange the current location between nearby devices
- **Database**: store and manage location information
- **Human Machine Interface**: display devices' location on map

B. Architecture design

To develop applications on the ITS station reference architecture [7], the first step is to identify which requirements should be inside application or other entities. The potential architecture designs are, therefore, as follows:

- **Application-based solution**: a traditional self-contained solution, in which each application contains all the necessary functions inside itself. Applications do not use the Facilities layer and management entity functions but, directly use the networking and transport layer features.
- **Facility-based solution**: most of the common functions are integrated into the Facilities layer and Management entity, while applications simply use them. In this solution, the discovery, communication, and database functions can be supported by ongoing ITS standards: *Generic message distribution handler*, *Communication profile handler / System management entity*, and *Local dynamic map*, respectively [10].

Although the application-based solution is simple, it lacks interoperability, extensibility, and incurs the cost of communication handling; it is inefficient that each application has redundant features, especially in hardware with limited capability. Furthermore, as the ITS station reference architecture accepts multiple access technologies and communication protocols, it is burdensome to support all of them in each application. A solution is to static configuration: exclusively use a certain type of protocol, however, it prevents to adapt dynamic mobile networks.

On the other hand, in the latter solution, most common functions are supported in the Facilities layer. Redundant development are minimized so that applications can concentrate on their specific task. Moreover, applications do not need to take care the diversity of the access technologies and communication protocols, because it is also handled by facilities layer services. Thanks to its efficiency of development and the adaptability of ITS standards, we adopt the facility layer solution to implement the proposed application, as depicted in Figure 2. The following sections detail the relevant components and their interaction.

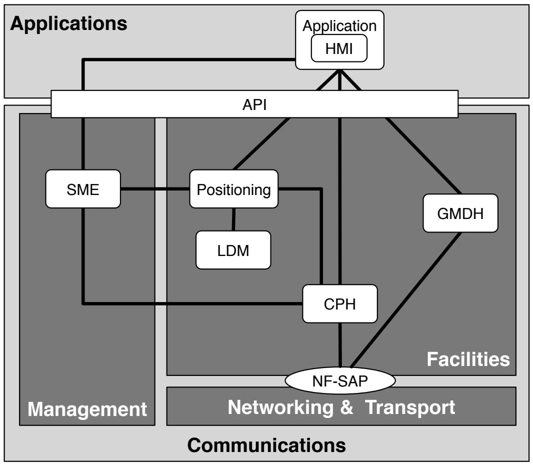


Fig. 2. Selected system architecture in compliance with the ITS-S reference architecture with the 5 conceptual components: Positioning, Communication Profile Handler (CPH), Generic Message Distribution Handler (GMDH), System Management Entity (SME) and the Local Dynamic Map (LDM).

C. Application

The application is composed of its main logic and Human Machine Interface; for this reason, it belongs the *Application* horizontal entity from ISO 21217. When the application is started, it requests position information to *Positioning* entity. Once the position is received, the application shows it on a graphical user interface (map) and communicates with *Communication profile handler* to send the Application Data Unit, which contains device's current position, timestamp, accuracy and its identifier.

D. Facilities and Management

As described previously, we implement five conceptual components in the Facilities layer and Management entity: *Positioning, Communication Profile Handler, Generic Message Distribution Handler, System Management Entity*, and *Local Dynamic Map*:

Positioning is an entity to process the device's position information. This entity provides application and information supports, for this reason it belongs to the Facilities layer. It complies to the ITS standards as a data source for *Local dynamic map* and *Application*, abstracting this entities from the diversity of the source of position (GPS, static configuration file, CAN Network, Internet, etc.). This way, these entities do not need to take care each data sources.

Communication Profile Handler (CPH) based on the ISO 17429 [10], it enables applications to abstract the diversity of communications. With this component, applications can transparently use multiple communication protocols and access technologies. Only applications need to do is to register their *communication requirements*, the type of communication, destination, quality, priority, etc. This communication requirement is mapped to each application, and then *Communication* *profile handler* configures the underlying communication stack to satisfy the requirement.

Generic Message Distribution Handler (GMDH), based on the ISO 17429 [10], it is used to share a specific message among multiple applications by means of the publish/subscribe scheme. This scheme enables the push-based commutation, in which each receiver application *subscribes* a certain type of message while sender application *publishes* any message regardless of the presence of receiver. The message is delivered only when there are applications that subscribe to this message. We implement *Generic message distribution handler* to exploit its publish/subscribe mechanism for supporting information sharing and discovery.

System Management Entity (SME) this entity can be used by all entities. It enables cross-layer services by storing information from any communication layers [7]. In this paper, we implement *System management entity* to manage the communication profile as described by ISO 17423 [8].

Local Dynamic Map (LDM) is a database containing static maps, static information not yet part of the above maps, temporary and dynamic information and dynamic information concerning moving objects as defined by ETSI [12]. In this paper, we implement *Local dynamic map* as a data store of the position information, which can be requested by other applications. As described above, the position information is provided by the Positioning entity.

E. Interaction model

The proposed application and facilities communicate in compliance with the following operational steps: *flow assignment*, *position management*, *message transmission*, *message reception*, and *user interaction*.

In the flow assignment operation, the application presents its communication requirements to *System management entity*, then it generates and returns a *FlowTypeID*, an identifier to map the application to communication requirement [8] [13]. When the application wants to send messages, it registers the destination with previously-assigned FlowTypeID to *Communication profile handler*, which configure an appropriate communication stack and generates/replies a *FlowID* (an identifier of communication flow mapped to the application), its communication requirement, and destination. To transmits messages, the application passes message body to *Communication profile handler* with FlowID. The flow assignment is depicted in Figure 3.

To perform the position management operation, the application at first establishes connection with the *Positioning* entity, and then obtain the device's current position at any time. The *Positioning* entity also performs the flow type registration and flow registration as an application, because application may access the positioning entity in the remote host.

Once the *Positioning* is started, it communicates with *Local dynamic map* to search possible position information previously stored. If there is no position information, *Positioning* communicates with GPS to get them and store this position on the *Local dynamic map*. In this paper, *Positioning* and *Local dynamic map* were specified as a pair of separate functions.

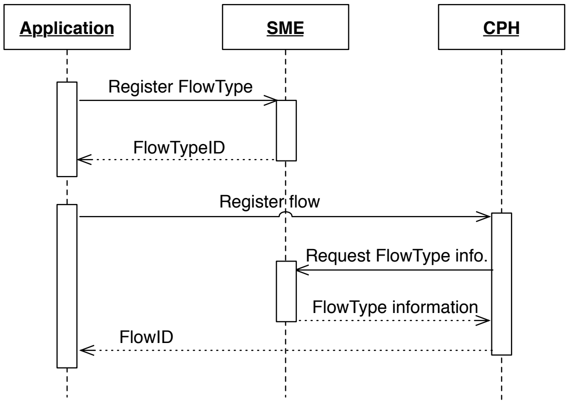


Fig. 3. Flow Assignment.

Once FlowID is assigned, the application sends its current position obtained from the above operation. In contrast to using traditional socket APIs, the application passes *Application Data Unit* (ADU) to *Communication Profile Handler* (CPH) with FlowID, then it publishes *Application data unit* to a specific destination. Figure 4 shows the message transmission operation. For the proposed application, the destination is single-hop broadcast.

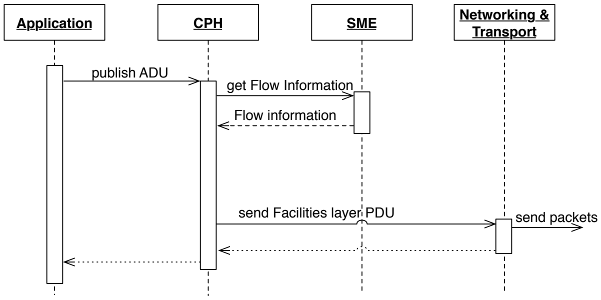


Fig. 4. Application Data Unit Transmission.

To receive location information, each application *subscribes* this message to *Generic message distribution handler*, then it distributes the message to applications only when the corresponding message is received. This communication follows the passive, push-based manner.

The application interacts with users via human machine interface to display its own location and the nearby mobile devices' location.

IV. Implementation

All the components are implemented using Android Software Development Kit (SDK), targeted to Android 4.0 or later. The application is a set of foreground Android activities, while the Facilities layer and Management entity components are *Android services*, application-independent background processes. To get position information, device's built-in GPS is used via Google play services library. As a Human-machine interface, we use Google Map. Wi-Fi direct [14] is used for device-to-device direct communication.

A. Boundary of entities

Local dynamic map and Positioning are implemented as two independent Android services, while *System management entity*, *Communication profile handler* and *Generic message distribution handler* are a single service; because these three entities are dedicated for communication, we coupled them for performance reason (data can be more directly accessed among the entities). Note that such a decision, i.e., coupling the conceptual entities, does not affect the compliance to the standards: the definition of the entities in the standards are conceptual, therefore, how to couple the functions are developers' choice as long as it is compliant to the standardized interfaces.

Since each entity is stand-alone process, the interaction among the application and each component is performed as Inter-process-communication using Android Interface Definition Language (AIDL) [15]. This way, each facility provides APIs to uses as AIDL interface file; then, the users transparently use the provided functions via APIs without considering the inter-process-communication.

B. Communication

Although the traditional way to share information between devices is indirect communication using centralized hosts, we chose the device-to-device direct communication without servers because the intended scenario is transient communication among nearby pedestrians (mobile devices) which only requires single-hop broadcast. For this reason, Android's built-in Wi-Fi direct (called *Wi-Fi Peer-to-Peer* in Android), which enables to discover and connect to other devices, is used. The manipulation of Wi-Fi direct is implemented in the *Communication profile handler/Generic message distribution handler* service.

Regarding the flow assignment, in the current implementation, we introduce some experimental *well-known* FlowTypes: statically configured communication requirements stored in the local storage, i.e., the type of transport layer protocols and source/destination address and port number. These static settings are loaded when the *Communication profile handler* service is started, and then users specify one of the requirement by FlowTypeID (assuming well-known FlowTypes are publicly available a priori).

C. Application Programming Interface (API)

Each component provides a number of APIs to interact with applications. *Positioning* provides

```
messengerToServicePositioning.send(msg)
```

which returns the current position, where `msg` is an Android's *Message* object for inter-process communication, which contains description of the request from applications, such as the type of request and application's Identifier.

To manage position, *Local Dynamic Map* provides

```
messengerToServiceLDM.send(msgToLDM)
```

which stores or returns position requested by *Positioning* entity, where `msgToLDM` is an object containing request description, such as the type of request and ITS station's Identifier.

On the other hand, to send *Application data unit*, *Communication Profile Handler* provides:

```
publish(int flowId, List messageIds,
byte[] adu)
```

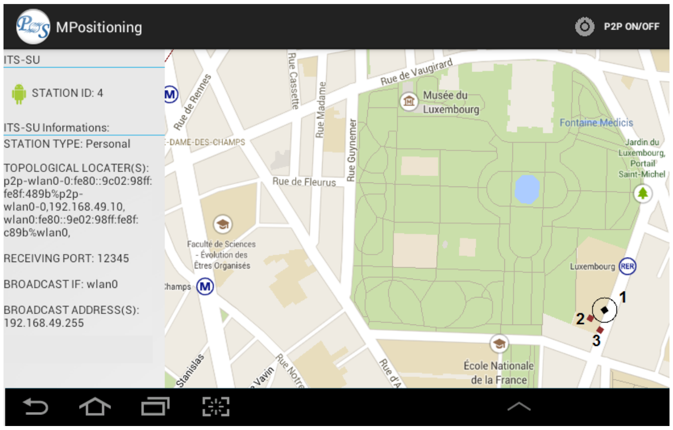
where `flowId` is an identifier of a destination mapped with communication requirements, `messageIds` and `adu` is the type and serialized sequence of bytes of *Application data unit*. How to encode and decode `adu` is application's responsibility.

D. Message format

The *Application data unit* object is composed of `latitude`, `longitude`, `accuracy`, `timestamps` and `stationID` attributes. To efficiently exchange data between devices, *Application data unit* is formatted, encoded and decoded according to ASN.1. We use *BinaryNotes* [16], an open source ASN.1 framework, to encode and decode the *Application data unit*.

E. Initial demonstration

We installed the application and services into three Android tablets (Samsung Galaxy Tab 10.1 GT-P5100, Android 4.1.2). Initial demonstrations have been performed using these devices, and shown each device properly exchanges its location information. Figure 5 shows the screenshot of the application displaying the location of neighboring devices.



ITS-SU

STATION ID: 4

ITS-SU Informations:
STATION TYPE: Personal

TOPOLOGICAL LOCATER(S):
p2p-wlan0-0:fe80::9c02:98ff:fe8f:489b%p2p-wlan0-0,192.168.49.10,
wlan0:fe80::9e02:98ff:fe8f:c89b%wlan0,

RECEIVING PORT: 12345

BROADCAST IF: wlan0

BROADCAST ADDRESS(S):
192.168.49.255

Fig. 5. Application's user interface displaying two neighbors.

In Figure 5, the device's own location is identified by *Station ID 1*, while the nearby nodes' locations are *2* and *3*, respectively.

V. Discussion

This paper demonstrated how conceptual entities defined by ITS standards can be implemented. As the ITS standards describe a framework to develop ITS applications, in this paper, we proposed a simple application to evaluate how applications and the underlying functions can be implemented, and also their interactions. This section describes practical consideration of the implementation of ITS standard, and issues of ITS applications on Android.

Because standards describe minimum sets of essential features for interoperability, we have studied a number of design choices. A main choice is how to integrate a wide variety of conceptual elements in the standards into actual implementation. A simple solution is to make a single self-contained software component, while the other way is to actually separate each entities. In other words, it is the choice of single binary or multiple modules. In general, the former solution is superior in terms of performance: if we couple all conceptual entities into a single component, the interface between the entities are much simpler. However, this solution lacks extensibility and is difficult to maintain, specifically if it is developed by multiple stakeholders. In other words, the modular solution is efficient in terms of interoperability. The important design choice in this solution is the granularity of modules: implementing each conceptual module to exactly one component may need redundant interaction. The number of stakeholders and capability of target devices, therefore, should be carefully considered.

We used Android's WI-FI Direct for device-to-device direct communication; however, during the demonstration, we observed weaknesses of this technology: whenever a device is detected by other Wi-Fi Direct enabled devices, it requires users' interaction (a confirmation box is pop-upped and users need to tap the button to accept connection for each device). Although it is secure to prevent unwanted silent connections from unknown devices, it cannot be used by ITS applications which needs quick and automatic/silent connection establishment. It is necessary to investigate the way of secure ad-hoc communication without users' interaction.

In this paper, our first application did not concern privacy issues as specified by [7]. Since, identity information, such as a pair of device/application/user identifier and its position, should not be unnecessarily broadcasted over the air; as a next step it is necessary to integrate ITS security related functions e.g., authentication and encryption.

VI. Conclusion and Future Work

Cooperative Intelligent Transportation Systems is an increasingly important topic to enhance our stressed transportation systems and address some safety and efficiency problems. Based on the Internet OSI layered model, transportation and communication communities are designing a modular architecture intended to be deployed soon. Standards are being written and this will hopefully ensure interoperability of very different systems. The goal of this paper is to share the knowledge we got from the design and implementation of a simple application compliant with theses new ITS standards. We have shown that there are some challenges in the organization of the modules and the concepts associated. Since this is true for a very basic application, we expect more problems when real applications will be implemented, especially for safety critical applications. There is a clear need for clarification of the concepts and module to be used and this papers intends to pave the way in that direction.

However, these difficulties should not elude the most important result of our work: it is possible — and finally, not that difficult, if carefully handled — to implement a Cooperative ITS application using the best of the modular approach described in the standards. This validates the standardization effort. The direction is promising and need more exploration. Some directions are clearly shown by this paper: imagine a set of applications sharing the same services of the Facilities layer; refine conceptually the 5 components we introduced so that more applications can exploit them maybe introducing additional components; tackle the issues of the platform functions needed for ITS applications (Android had some drawbacks; but, is a platform to address, at least for pedestrians). Moreover other functions provided by standards are to be carefully linked to ITS application: encryption, privacy, etc. We hope to see in the near future more and more works describing how to best deploy the promising Cooperative ITS architecture.

References

[1] R. Michoud, A. M. Orozco, and G. Llano, "Mobile ad-hoc routing protocols survey for the design of VANET applications," Intelligent Transportation Systems Symposium (CITSS), 2012 IEEE Colombian, pp. 1 – 6, 2012.

[2] P. Muhlethaler, Y. Toor, A. Laouiti, and A. de La Fortelle, "Vehicle ad hoc networks: applications and related technical issues," IEEE Communications Surveys and Tutorials, vol. 10, no. 3, pp. 74–88, Quarter 2008, URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?isnumber=4625798&arnumber=4625806&count=7&index=6 [accessed: June 2014].

[3] P. Papadimitratos, A. de La Fortelle, K. Evenssen, R. Brignolo, and S. Cosenza, "Vehicular communication systems: Enabling technologies, applications, and future outlook on intelligent transportation," Communications Magazine, IEEE, vol. 47, no. 11, pp. 84–95, November 2009.

[4] "IPv6 ITS Station Stack (ITSSv6) European project," URL: https://project.inria.fr/itssv6/ [accessed: June 2014].

[5] "Drive C2X European project," URL: http://www.drive-c2x.eu [accessed: June 2014].

[6] "Cooperative Vehicle-Infrastructure Systems (CVIS) European project," URL: http://www.cvisproject.org/ [accessed: June 2014].

[7] "ISO 21217:2014 Intelligent transport systems - Communications access for land mobiles (CALM) - Architecture," March 2014.

[8] "ISO/DTS 17423 Intelligent transport systems - Cooperative systems - ITS application requirements for automatic selection of communication interfaces," February 2013.

[9] "ISO/DTS 17419 Intelligent transport systems - Cooperative systems - Classification and management of ITS applications in a global context," February 2013.

[10] "ISO/NP 17429 Intelligent transport systems - Cooperative systems - Profiles for processing and transfer of information between ITS stations for applications related to transport infrastructure management, control and guidance," December 2012.

[11] "ISO 24102-3:2013 Intelligent transport systems - Communications access for land mobiles (CALM) - ITS station management - Part 3: Service access points," June 2013.

[12] "ETSI TR 102 893 V.1.1.1 Intelligent Transport Systems (ITS) - Security - Threat, Vulnerability and Risk Analysis (TVRA)," March 2010.

[13] "ISO 24102-1:2013 Intelligent transport systems - Communications access for land mobiles (CALM) - ITS station management - Part 1: Local management," June 2013.

[14] Wi-Fi Alliance, "Wi-Fi Direct," URL: http://www.wi-fi.org/discover-wi-fi/wi-fi-direct [accessed: May 2014].

[15] "Android Interface Definition Language (AIDL)," URL: http://developer.android.com/guide/components/aidl.html [accessed: June 2014].

[16] "BinaryNotes :: ASN.1 framework," URL: http://bnotes.sourceforge.net/ [accessed: May 2014].

Extending a 3GPP Prepaid Protocol to Improve Credit Pre-reservation Mechanism

Natal Vieira de Souza Neto,
Flávio de Oliveira Silva
and Pedro Frosi Rosa

Faculty of Computing
Federal University of Uberlândia
Uberlândia, MG, Brazil
Email: natal@mestrado.ufu.br,
flavio@facom.ufu.br,
pfrosi@ufu.br

Marcos Guimarães de Medeiros
and João Henrique de Souza Pereira

Innovation, Research and Development
Algar Telecom
Uberlândia, MG, Brazil
Email: marcosgm@algartelecom.com.br,
joaohs@algartelecom.com.br

Abstract—**Online Charging System (OCS) uses credit pre-reservation to rating prepaid services. As the current standardized protocols present some problem during the pre-reservation phase, a new service must be introduced to address this kind of problem. In this paper, we propose a protocol for the communication between OCS modules. Running at the application layer, the protocol focuses on the definition of Refuse message, by regarding the credit reservation for call service, to be used when Rating Function module has some performance problem. We formalize the protocol messages and present its design rules and procedures. As results, the protocol was implemented into production in a carrier, and all calls were completed after the new feature.**

Keywords–Credit reservation; Telecommunication protocol; Online charging; Prepaid service.

I. Introduction

Online Charging System (OCS) [1] is a platform for charging, rating and controlling calls and other services like Short Message Service (SMS), Multimedia Message Service (MMS), General Packet Radio Service (GPRS) and Value-Added Service (VAS). In a good architecture for telecommunication enterprise, all prepaid services are charged by OCS. Prepaid service is the service for which the payment is made before the use [1].

In typical OCS, as recommended by 3rd Generation Partnership Project (3GPP) technical specifications, the module of call control must be separated from the module which processes the specified rules. The main reason is to avoid the overhead for the the processing module [1]. The module for call control is named Session Based Charging Function (SBCF) or Event Based Charging Function (EBCF). For the purpose of this work, we use the name Based Charging Function (BCF), because the most important here is the communication protocol and not the module development. The module for processing pricing and prepaid rules is named Rating Function. Call control with other application platforms can be built on BCF. Prepaid rules, pricing rules, Charging Detail Record (CDR) [2] records and database interaction can be built on Rating Function.

The communication between BCF and Rating Function must be made in real time, as the session is spread into two modules. A protocol is necessary for the communication to keep the call. This protocol is described in [3], but it does not address the Rating Function performance problem. Regarding the performance, a specific primitive is required, indicating to Rating Function that a problem happened and has been handled. This paper proposes an extension to the 3GPP protocol to address the issues described above, introducing the Refuse primitive, that besides the protocol specified here, could be used in other protocols for unit reservation, for instance.

A protocol that solves the problem of Rating Function and BCF communication is a protocol which ensures the session of the call on BCF and Rating Function. When a network element on telecommunication architecture, for instance Mobile Switching Center (MSC), communicates with OCS, the protocol used for mobile call is Camel Application Part (CAP), which initializes and finalizes a call. Newer architectures, for instance IP Multimedia Subsystem (IMS), use Diameter protocol and the communication is made by Diameter Credit Control (DCC). The CAP or DCC communicates with BCF, but the communication between BCF or Rating Function normally is made by Diameter [4], or proprietary protocols.

With a scale-out architecture using commodities hardware to build Rating Function, for instance, timeouts can occur when the call is initialized. The communication with database or other module made by Rating Function can be delayed. In this scenario, the primitive proposed in this paper ensures error control, which is not provided by 3GPP Prepaid Protocol specification.

Several protocols have been used to address this issue, but no one, as far as we know, has been successfully applied when subscriber's refund is necessary due to production environment. Based on it, a new service named Refuse is proposed. This paper focuses on call service, and not on GPRS, SMS, MMS or VAS.

The remaining of the paper is structured as follows: in Section II, we present related works; in Section III, we detail the credit pre-reservation mechanism; in Section IV, we describe our protocol specification and design, i.e., its environment, encoding, vocabulary, services and procedure rules; in Section V, we present the implementation details behind our approach; and finally, in Section VI, we make our final remarks.

ISBN: 978-1-61208-360-5

II. Related Work

Credit pre-reservation is not a new ideia. According to 3GPP TS 32296 [1], an EBCF requests Rating Function for unit reservation. Unit reservation and credit pre-reservation are similar ideas.

When a call or other service is initialized, OCS works while the call is on. The performance is analyzed in others papers, like [5]. It means that OCS performance is a feature to be carefully analyzed. The Refuse primitive is necessary when the pre-reservation phase presents some problem.

Several services that are controlled by the OCS need credit pre-reservation. One example is Universal Mobile Telecommunications System (UMTS) prepaid service, exposed in [6]. In this scenario, the DCC protocol provides pre-reservation. To facilitate the understanding, some primitives in this paper are similar to those from DCC. We introduce the INIT_CALL_FIXED and INIT_CALL_MOBILE, similar to DCC INITIAL_REQUEST message. The DCC TERMINATE_REQUEST is similar to FINISH message described here. A telecommunication platform, like General Packet Radio Service (GGSN), can communicate with OCS using DCC. Into the OCS, BCF will communicate with Rating Function or other modules, and this communication uses the services INITIAL_REQUEST and TERMINATE_REQUEST. Then, BCF uses primitives presented here to keep this flow inside OCS. This paper introduces Refuse service to assure that the process will finish properly.

IMS aims to evolve network core from circuit-switching to packet-switching. IMS, standardized by 3GPP, becomes real in many networks. OCS is an element embedding in an IMS architecture. Communication in IMS often uses Diameter, and DCC, as described previously, has a flow for the credit reservation. The protocol proposed here is used inside OCS [7].

The 3GPP has standardized it by regarding flexibility and to provide all telecommunications functions over IP. OCS is an element embedding in an IMS architecture. Communication in IMS often uses Diameter, and DCC, as described previously, has a flow for the credit reservation. The protocol proposed here is used inside OCS [7].

A single module used for rating function was designed by Oumina and Ranc [8]. Rating Function integrates complex applications and allows the specification of quality of service parameters. Since it is a real time system, timeouts between Rating Function and BCF could happen.

III. Credit pre-reservation mechanism

The communication with OCS platform is made in telecommunications architecture based on real time systems. When a subscriber makes a call, the session request is sent to OCS, which returns complete or not complete. There are two types of terminals: fixed (wired) or mobile (wireless). The fixed architecture, for standard, uses Session Initiation Protocol (SIP) [9], and mobile architecture uses CAP protocol [10]. Furthermore, modern architectures use Diameter protocol. The basic difference between the three protocols is the way they establish the session. SIP sends the call session request to OCS, while CAP maintains the session on MSC, and only triggers OCS. The subscriber's money amount is controlled by OCS, but the session is controlled by MSC. The diameter uses DCC to trigger the OCS, and DCC controls the session [11][12].

In all cases described, the trigger or session come on BCF, which will control the session or just receive the trigger. If it receives a call attempt, it could make a numerical analysis, for instance, push an insertion network of number A or number B, or others [13]. Then, it sends the call informations to Rating Function, where the charging processing is made. When Rating Function receives a call attempt, it will analyze if the number can (or cannot) make the call, which is named Init. The matter here is the time. When MSC (or other telecommunication element) initializes a call, it needs the time that call will last. As the call is prepaid, the charging must be made at this moment. However, it is highly network consuming to trigger the Rating Function every second, thus Rating Function must decide the time granted for the call [3].

After the call, the BCF communicates to the Rating Function the duration of the call. If the duration of the call is less than the granted time, the Rating Function returns the difference. In the previous example, the time reserved was five minutes, supposing the call duration was two minutes, Rating Function returns three minutes to be refunded to the subscriber [8].

However, in some cases, the call is established and its duration is greater than the granted time. For instance, if Init granted five minutes, but the call took six minutes, the BCF will ask for more time. This is done through a service named Continue Call. Mandatorily, Init and Finish services are requested once, but Continue Call could be requested zero, one or more times [12]. The described mechanism is the credit pre-reservation.

IV. Protocol Specification and Design

For the protocol specified in [1], the pre-reservation mechanism will take place on BCF modules. In this design, the protocol will reserve a balance on Rating Function because performance evaluation shows that doing it on BCF module takes more time than doing it on Rating Function. As BCF should always complete the calls, independently of Rating Function time settings, all rules should be implemented on the Rating Function. Furthermore, the name of the proposed services also changes, because the communication flows are changed so that only the Rating Function will handle the balance.

The basic elements to specify a protocol are: the services provided by the protocol, the assumptions about the environment where the protocol will be executed, the message vocabulary used, message format and procedure rules that ensures consistence [14].

With the specification, it is possible to validate the protocol. This is done by creating the validation model, where the following features will be modeled: timings, flow control, message channel, process type, variables and data types, state execution, procedure model and recursion. After validation model, there are correctness requirements, which define the behavior, assertion, deadlocks, and so on. After the correctness requirements are defined, we are ready to design the protocol [14].

For the protocol proposed in this paper, the client is the BCF and the server is the Rating Function. It is a typical client-server relationship.

The assumptions about the environment start by considering that all of the actions (for instance credit reservation) should be completed before the call establishment. As the telecommunication industry is highly regulated, many of these actions should happen within a time interval.

As protocols are essentially event driven (services), it is important to specify what are the names of these services, in our case, there are nine services.

A. INIT_CALL_FIXED

this service is invoked when a call is started from a fixed terminal. BCF receives an indication to initialize a call, typically through SIP protocol. Then, BCF sends this service request to Rating Function, indicating that the call control is waiting for a response from the Rating Function whether this call can (positive) or cannot (negative) be completed.

B. INIT_CALL_MOBILE

this is similar to the previous service, but for a mobile terminal. It is necessary because there are different parameters which are sent when the terminal is mobile, like International Mobile Subscriber Identity (IMSI).

C. INIT_CALL_RESPONSE

upon INIT_CALL_FIXED or INIT_CALL_MOBILE indication, Rating Function processes all of the rules are applied to the requester terminal and returns to BCF the Call Id and the granted time, if BCF successfully completes the call. Otherwise, a negative answer is returned containing the error code.

D. CONTINUE_CALL

if the time granted is not enough to support the call, BCF can ask for additional time by requesting CONTINUE_CALL service informing the Call Id. Upon its receive, the Rating Function applies the charging rules and decides to grant, or not, a new time interval for the call.

E. CONTINUE_CALL_RESPONSE

upon receiving a CONTINUE_CALL, Rating Function processes the request, all of the rules are applied to the requesting Call Id, and, if successful, a new granted conversation period is returned to BCF for this Call Id. The call is ended if the granted period is zero.

F. FINISH_CALL

this service is invoked when the call is terminated by subscriber's terminal or the time is insufficient to keep the call.

G. FINISH_CALL_RESPONSE

it informs to BCF that the call has been correctly finished and, thus, its session removed and the balance updated. As it is a response, if timeout occurs, BCF generates a CDR ERROR, indicating the charging has been made incorrectly.

H. INIT_REFUSE_RESPONSE

this message has been introduced in this work to address performance. When BCF sends INIT_CALL_FIXED or INIT_CALL_MOBILE, it waits for INIT_CALL_RESPONSE. Depending on timing parameters configured in other telecommunications elements (like MSC), if there is a delay in the response, affecting other elements, it could disturb the Finite State Machine (FSM) in all of the telecommunication architecture elements. We implemented INIT_REFUSE_RESPONSE to address this issue. If timeout occurs between BCF and Rating Function, BCF completes the call (in an offline manner), independently of the Rating Function's response. However, due to the asynchronous behavior of the network, BCF could receive, after timeout, the previously expected INIT_CALL_RESPONSE. As BCF initialized an offline call, it sends to Rating Function an INIT_REFUSE_RESPONSE, instructing Rating Function to refund customer the pre-reserved credit.

I. FINISH_REFUSE_RESPONSE

this message has been introduced in this work to improve the call error control. When BCF sends FINISH_CALL, it waits for a FINISH_CALL_RESPONSE, indicating that Rating Function correctly updated the subscriber's credit. BCF sends to Rating Function a FINISH_CALL, and if the timeout occurs, BCF considers that Rating Function failed, and generates a CDR ERROR, ie, a CDR indicating that the call could not be finished properly, and it sends to Rating Function a FINISH_REFUSE_RESPONSE, informing the Rating Function that call may have had charging problems.

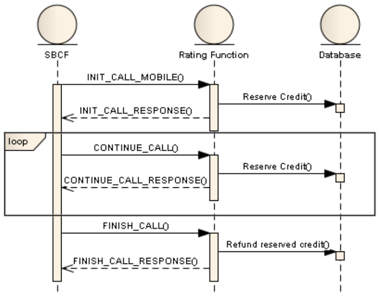


Figure 1. Call flow.

As shown in Fig. 1, BCF sends INIT_CALL_MOBILE (or INIT_CALL_FIXED), upon receiving a call request from other telecommunication elements. BCF is responsible for the call control for mobile calls, and maintains the session control for fixed calls. In fixed calls, BCF counts the duration, but for mobile calls, it is made by other elements, because CAP has parameters informing BCF of the duration of the call. BCF sends to Rating Function the Init, an Init for each call, independent the calling number. In modern platforms, a single calling number can make several calls at a time. Upon receiving INIT_CALL_MOBILE, the Rating Function creates a session with the rating information.

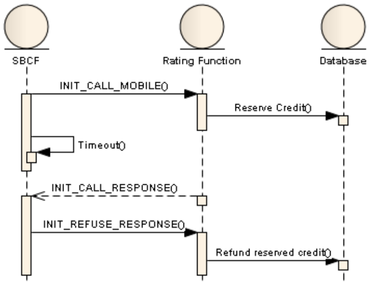


Figure 2. Call flow with Init Refuse.

In Fig. 1, upon receiving INIT_CALL_MOBILE, the Rating Function returns an INIT_CALL_RESPONSE which contains the duration for the call and, if needed, BCF could

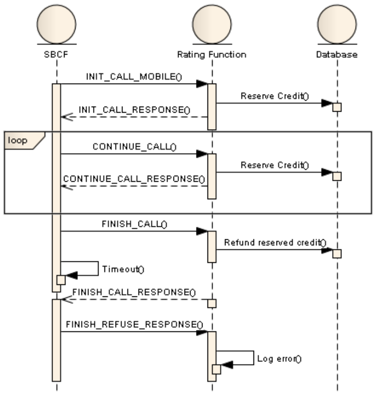


Figure 3. Finish Refuse service behavior.

ask for more time. If the duration is finishing, BCF requests a CONTINUE_CALL service for the particular call id. In architectures with modules installed in several machines, a CONTINUE_CALL must be sent to the specific Rating Function which received the INIT_CALL_MOBILE. Then, Rating Function returns more duration or not continues the call. Several CONTINUE_CALL messages can be issued, until the subscriber ends the call, or there is no more money in the account.

Fig. 1 also shows the FINISH_CALL service, where subscriber ends the call, or the subscriber has insufficient balance. BCF requests FINISH_CALL by indicating the call id and the call duration. With these pieces of information, the Rating Function can finish the call, crediting money (if the credit reservation was not fully used), and generating CDR.

When some problem happens at the beginning of the call, the flow is similar to Fig. 2, when BCF receives a request to initialize a call, and starts the call control. BCF sends INIT_CALL_MOBILE (or INIT_CALL_FIXED), as shown in Fig. 2, some problem happens by regarding Rating Function. BCF starts an offline call and the customer will not be charged. An offline call is released with unlimited duration and BCF will wait for another platform (like MSC) to finish the call. Even delayed, the Rating Function will reserve credit and answer to the BCF, and BCF will receive after the timeout. As BCF started an offline call, Rating Function response is not applicable. Upon this, BCF sends an INIT_REFUSE_RESPONSE telling Rating Function to refund subscriber the reserved credit. At the end of the call, BCF generates an offline CDR which will be eventually manually tariffed.

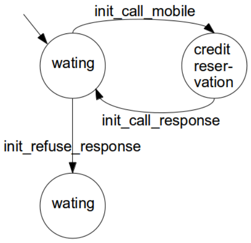


Figure 4. Rating Function Finite State Machine for Init Call service.

Fig. 3 represents the FINISH_REFUSE_RESPONSE which is used when a FINISH_CALL_RESPONSE delays or does not come to BCF. BCF informs the Rating Function that the call is facing a charging problem and the call is logged for future manual analysis. When FINISH_REFUSE_RESPONSE is used, BCF generates a CDR ERROR, for further analysis.

Now, we will define the services. In this paper, three services will be analyzed: Init, Continue and Finish. The Continue service does not have refuse message, as described previously.

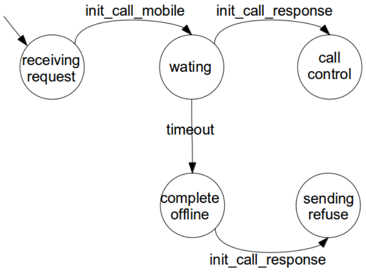


Figure 5. BCF Finite State Machine for Init service.

The Init service is shown in Fig. 4 which represents the FSM of Rating Function. Basically, Rating Function waits INIT_CALL_FIXED or INIT_CALL_MOBILE (as shown) to proceed with the credit reservation and responses with a INIT_CALL_RESPONSE message. If Rating Function receives an INIT_REFUSE_RESPONSE, it must refund the pre-reserved money. Fig. 5 represents Init service at BCF. In the first state, BCF is waiting a request from customer terminal (Calling User). Upon receiving it, it initializes the call control, sends to Rating Function a INIT_CALL_MOBILE and waits for a response. If it receives the INIT_CALL_RESPONSE primitive, the call is completed and becomes under control. However, if timeout occurs, BCF completes the call in offline mode, and if INIT_CALL_RESPONSE comes after timeout, the INIT_REFUSE_RESPONSE primitive is sent to Rating Function.

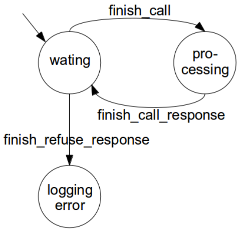


Figure 6. Rating Function Finite State Machine for finish service.

The Finish service at Rating Function is represented in Fig. 6. First, the Rating Function is waiting for the end of the call. When the call ends, the Rating Function receives a FINISH_CALL primitive and processes it (updating subscriber's accounts and generating CDR). However, if Rating Function receives a FINISH_REFUSE_RESPONSE, it logs the information, as it is impossible to know if the call was correctly or incorrectly charged.

Fig. 7 shows the FSM at BCF, where if a timeout on finish service happens, a FINISH_REFUSE_RESPONSE is sent to Rating Function. A CDR ERROR is generated which will be manually treated later.

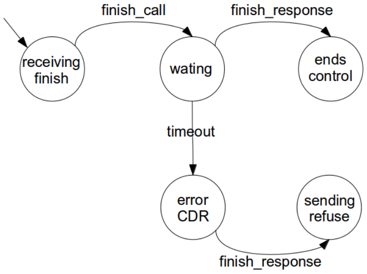


Figure 7. BCF Finite State Machine for Finish service.

V. Implementation Details

The first development of the protocol, to build BCF Module, was made using Java language and JAIN SLEE Mobicents [15]. The Rating Function Module also was built in Java, by using Spring Framework [16]. The main reasons for this are due to the fact that Mobicents has an architecture designed to create, deploy and manage services and applications integrating voice, and other services. Mobicents has been largely used for telecommunication companies around the world. Spring allows several programming techniques, which is helpful because the Rating Function has several rules. All rules were tested and deployed in a real Brazilian telecommunications company.

The formatting for the protocol communication is made by using Protocol Buffers (protobuf). It is a well tested framework for the encoding of structured data [17]. The primitives were packed/unpacked using message concept of protobuf which allows the focus on the design and implementation of the protocol. Protobuf converts the defined messages to Java classes which can be used on BCF (with mobicents) or Rating Function (with Spring). The messages contain typed fields. Each message will be described below.

A. INIT_CALL_FIXED message fields

1) Call Id: this parameter is a string that contains the session's Id.

2) Insertion_Network: this paramenter is a string and for fixed calls, Next Generation Networks (NGN) could insert a prefix in calls, to distinguish some services. It may contain digits, characters and special symbols.

3) Original_Number_A: this parameter is a string and it contains the original calling number. It may contain digits, characters and special symbols.

4) Tariff_Number_A: this parameter has the same type as Original_Number_A which is modified by BCF before sending it to Rating Function.

5) Original_Number_B: this parameter is a string and it contains the original called number. It may contain digits, characters and special symbols.

6) Tariff_Number_B: this parameter has the same type as Original_Number_B which is modified by BCF before sending it to Rating Function.

7) CSP: this paramenter is a string and it contains the Carrier Service Prefix, used to make long-distance calls. It may contain digits, characters and special symbols.

8) Flow: this paramenter is integer, and specifies the call flow, indicating if the BCF received a calling or a called number.

9) Call_Type: a string, indicating if the call is local (same region) or not (other region).

10)Flag_Portability_Number: this parameter is a boolean and it indicates if the calling or called number is a ported number (originally from other operators).

11)CNL_A: a string, indicates the locality of calling number. The price can be determined based on the distance between the calling number and called number, and this parameter is used for this.

12)CNL_B: a string, indicates the locality of called number, also used to determinate the distance between calling and called numbers.

13)Portability_Prefix: a string, indicating portability of the calling/called number. When a subscriber migrates of operator, but maintains the number, this parameter is needed to define to which carrier the number belongs to.

B. INIT_CALL_MOBILE message fields

This service has the same INIT_CALL_FIXED message fields, except CNL_A and CNL_B needed only by fixed number. Besides INIT_CALL_FIXED fields, INIT_CALL_MOBILE has the following exclusive fields:

1) Is_Video_Call: this parameter is a boolean and indicates if the call is a video call. CAP provides this information. In case of other protocols that BCF receives, the information does not exist. Therefore, is a optional field, and the default value is false.

2) IMSI: a string, indicating the IMSI. If the flow is calling, it is the calling number IMSI, if the flow is called, it is the called number IMSI.

3) Cell_Global_Id: a string, indicating the Cell Id. A cell is the sector of the Base Transceiver Station (BTS) or NodeB where the mobile is registered.

With the messages described above, Rating Function can handle and charge the call.

C. INIT_CALL_RESPONSE message fields

The INIT_CALL_RESPONSE is the same for both fixed and mobile calls. It is described below:

1) Call Id: a string, indicating the id, same id received by Rating Function on INIT_CALL_MOBILE or INIT_CALL_FIXED. It is necessary for the BCF to identify for which call it is receiving response.

2) Action: a string, it indicates to BCF if the call must be completed (or not), or if a ring tone must be played before answer.

3) Quantity: a long value, it is the duration released, ie, how long BCF must allow the call.

4) account: a vector of numerical types, with the accounts. Some money was reserved on these accounts. This field is important for BCF to maintain information about charging, for a possible refuse.

D. INIT_REFUSE_RESPONSE message fields

When necessary, BCF sends a Init Refuse service. The following fields are used:

1) Call Id: a string, indicating the id. It is necessary for the Rating Function to know which call was refused.

2) Action: a string, indicating the action received on INIT_CALL_RESPONSE. If the call was not completed, it will not be in Rating Function session.

3) Account: a vector of long, the account's vector received on INIT_CALL_RESPONSE. It is important because refuse indicates a problem in Rating Function, and maybe Rating Function has no call information. In this case, it returns balance for accounts on this vector.

E. CONTINUE_CALL message fields

Continue Call service, implemented with the CONTINUE_CALL message. It has the following fields:

1) Call Id: a string, the same id used on Init call. It is necessary for the Rating Function to know which call is requesting more time.

F. CONTINUE_CALL_RESPONSE message fields

The CONTINUE_CALL_RESPONSE message has only new quantity released and the accounts vector.

G. FINISH_CALL message fields

The Finish Call service is made with the FINISH_CALL message. This primitive has the following fields:

1) Call Id: a string, indicating the call id. It is necessary for the Rating Function to know which call has finished.

2) Start_Time: a date and time, indicating the start time of the call, this time is stored by BCF, and sent to Rating Function only on FINISH.

3) Total_Duration: a long value, in seconds, indicating the total duration, considering wait time of subscriber's equipment and the call duration.

4) Call_Duration: a long value, in seconds, indicating the total duration without the wait time.

5) Call_Charge: this is a boolean field and indicates if it is a charged (or not charged) call.

6) Finish_Cause: a string. Same protocols inform BCF of the finish cause, and BCF must send this information to the Rating Function. The Rating Function can use this information in CDR or in a specific rule.

H. FINISH_CALL_RESPONSE message fields

The FINISH_CALL_RESPONSE message has two fields: call id; and processed, a boolean field indicating if Rating Function finished correctly.

I. FINISH_REFUSE_RESPONSE message fields

The FINISH_REFUSE_RESPONSE message can be implemented with one unique field, the call id, indicating to Rating Function that a refuse happened.

All fields described above were used on a real development of an OCS platform. Depending on the rules used by OCS, new fields need to be added.

In tests in a real production environment, the rate of refused messages on the network is approximately 100 to 600 calls daily, in a universe of 4 million of calls daily. It is a small number, but refuse services are necessary so that the subscriber is not harmed.

VI. Concluding remarks and Future Work

In this work, a communication protocol between ECF (or SBCF) and Rating Function modules has been designed in OCS platforms. This protocol has introduced the Refuse services on the initiation/end of calls. The Refuse services can be used when the Rating Function delays answers or has problems, including being out of service. With these, the subscriber has its call established, without problem. The protocol was implemented and used, with heavy load tests.

We used Rating Function to reserve and to debit balance. The main reason for this was because the protocol proposed is similar to 3GPP protocol, but our design considers that the calls will be always completed. Then, SBCF or EBCF only make call control, becoming lightweight to control the call.

The result is a platform currently in use, which charges 900,000 subscribers, considering both fixed and mobile terminals, and controls Rating Function timeout, allowing calls to always be completed.

For future works, it is possible to see the same features provided through Refuse services to other platforms like GPRS, SMS, MMS and VAS. These services have different flows, and refuse messages must be designed considering other parameters and services.

References

[1] 3GPP, "Online Charging System (OCS): Applications and interfaces," 3rd Generation Partnership Project (3GPP), TS 32.296, Sep. 2011.

[2] ——, "Charging Data Record (CDR) parameter description," 3rd Generation Partnership Project (3GPP), TS 32.298, Dec. 2013.

[3] ——, "Charging architecture and principles," 3rd Generation Partnership Project (3GPP), TS 32.240, Dec. 2013.

[4] ——, "Diameter charging applications," 3rd Generation Partnership Project (3GPP), TS 32.299, Dec. 2013.

[5] T. Grgic and M. Matijasevic, "Performance metrics for context-based charging in 3gpp online charging system," in Telecommunications (ConTEL), 2013 12th International Conference on, June 2013, pp. 171–178.

[6] H.-Y. Lee and Y.-B. Lin, "Credit pre-reservation mechanism for umts prepaid service," Wireless Communications, IEEE Transactions on, vol. 9, no. 6, June 2010, pp. 1867–1873.

[7] H. Oumina and D. Ranc, "Specification of rating function of online charging system in 3gpp ip multimedia system (ims) environment," in New Technologies, Mobility and Security, 2008. NTMS '08., Nov 2008, pp. 1–5.

[8] ——, "Designing the rating function of 3gpp online charging system for ip multimedia subsystem," in Internet Multimedia Services Architecture and Applications, 2008. IMSAA 2008. 2nd International Conference on, Dec 2008, pp. 1–6.

[9] J. Rosenberg et al. SIP: Session Initiation Protocol. RFC 3261 (Proposed Standard). Internet Engineering Task Force. Updated by RFCs 3265, 3853, 4320, 4916, 5393, 5621, 5626, 5630, 5922, 5954, 6026, 6141, 6665, 6878. [Online]. Available: http://www.ietf.org/rfc/rfc3261.txt [retrieved: Jun., 2002]

[10] 3GPP, "CAMEL Application Part (CAP) specification," 3rd Generation Partnership Project (3GPP), TS 29.078, Dec. 2012.

[11] ——, "Advice of Charge (AoC) service," 3rd Generation Partnership Project (3GPP), TS 32.280, Dec. 2013.

[12] J. Bhosale and P. Pawar, "Credit pre-reservation mechanism for mobile prepaid service," International Journal of Innovations in Engineering and Technology, vol. 9, no. 6, Dec 2013, pp. 82–87.

[13] 3GPP, "IP Multimedia Subsystem (IMS) charging," 3rd Generation Partnership Project (3GPP), TS 32.260, Dec. 2013.

[14] G. J. Holzmann, Design and Validation of Computer Protocols. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1991.

[15] M. Femminella, R. Francescangeli, E. Maccherani, and L. Monacelli, "Implementation and performance analysis of advanced it services based on open source jain slee," in Proceedings of the 2011 IEEE 36th Conference on Local Computer Networks, ser. LCN '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 746–753.

[16] Spring. Springframework reference manual 3.1. [Online]. Available: http://spring.io/docs [retrieved: May, 2014]

[17] K. Varda. Protocol buffers. http://code.google.com/apis/protocolbuffers/. [Online]. Available: http://code.google.com/apis/protocolbuffers/ [retrieved: May, 2014]

Simulator of Multi-service Switching Networks with Multi-service Sources

Mariusz Głąbowski*, Dragana Krstić† and Maciej Sobieraj*
*Chair of Communication and Computer Networks
Faculty of Electronics and Telecommunications, Poznan University of Technology, Poland
Email: mariusz.glabowski@put.poznan.pl
†Department of Telecommunications, Faculty of Electronic Engineering, University of Niš, Serbia
Email: dragana.krstic@elfak.ni.ac.rs

***Abstract*—In this paper, we present a simulator of multi-service switching networks with various resource management mechanisms. The network can be offered three types of traffic streams: Erlang, Engset and Pascal, generated by multi-service sources. Each multi-service source can generate calls of a number of traffic classes. In addition, the paper presents an analysis of the influence of different parameters upon the traffic characteristics of switching networks.**

Keywords*–*multi-service sources; Erlang; Engset; Pascal; switching network; simulation.

I. Introduction

Working out effective and efficient methods for managing resources in nodes of communication networks, especially in the case of multi-service networks that rely on virtualization of resources, is a complex issue. One of the fundamental difficulties arises from the necessity of servicing by switching nodes of the networks different classes of traffic streams [1][2][3]. In order to obtain a desirable admission policy for calls of different traffic classes as well as a desirable level of usage of resources, many different strategies of resource management in multi-service network have been elaborated. The ones of the most effective strategies can be reservation mechanisms of resources, both dynamic (executed on-line and securing well-balanced access to network resources) [4], and static (executed appropriately ahead of time with time advancement) [4], as well as threshold [4][5] and priority mechanisms [4].

In order to fully determine the influence of resource management mechanisms on the effectiveness of switching nodes of telecommunications networks we have to determine their influence on switching networks that are the key element of each node in the network. In this paper we describe an original simulation program for a determination of traffic characteristics of switching networks with various resource management mechanisms and multi-service traffic sources.

The paper is structured as follows. Section 2 presents the structure of traffic offered to the multi-service switching network. Section 3 describes the control mechanisms used in switching networks to control access to resources. Section 4 includes a description of the simulator and presents methods for a simulation of different types of traffic (Erlang, Engset and Pascal traffic). Section 5 presents the results of the simulation experiments. Finally, the conclusions and summary of the paper are provided in Section 6.

II. Related work

The influence of resource management mechanisms on the effectiveness of multi-service communications systems can be determined on the basis of both analytical methods [6][7] and simulation tools [8]. In the case of the analytical methods, we can obtain a limited number of traffic characteristics, under limited assumptions. For example, none of methods of switching networks analysis, known to the authors, take into account other than random algorithms of link selection in outgoing directions or interstage links. So far, no analytical methods have been developed that allow switching network modeling with any given number of attempts to set up a new connection, and with traffic streams other than Erlang streams. The existing limitations of analytical methods have led to the development of a switching networks simulator [8]. The simulation model of switching network with traffic management mechanisms developed in [8], however, was limited to multi-service switching networks with single-service sources [8][9][10]. Consequently, in order to determine the influence of many parameters not taken into account in existing analytical models, and, additionally, the multi-service sources (a multi-service traffic source is the source that can generate calls of different classes [6][11]), the simulation model of the multi-service switching networks with multi-service sources is presented in the paper. According to the authors' knowledge, this is the first simulator of the switching networks with multi-service traffic sources.

III. Structure of offered traffic

In the considered model, m traffic classes that belonged to the set $\mathbb{M} = \{1, 2, ..., m\}$ were defined. A given class c is defined by the number t_c of demanded BBUs (Basic Bandwidth Units) required to set up a new connection of class c and the parameter μ_c of the exponential distribution of the service time of calls of class c [11]. The switching network is offered three types of traffic streams - Erlang, Engset and Pascal traffic streams. Each stream is generated by a source that belongs to an appropriate set of traffic sources: $\mathbb{Z}_{\text{Er},i}$, $\mathbb{Z}_{\text{En},j}$ and $\mathbb{Z}_{\text{Pa},k}$. In the considered system, s_I sets of traffic sources generating Erlang traffic streams, s_J sets of traffic sources generating Engset traffic streams and s_K sets of traffic sources generating Pascal traffic streams were defined. The total number of the sets of traffic sources in the system is $S = s_I + s_J + s_K$. The sources that belong to the set $\mathbb{Z}_{\text{Er},i}$ can generate Erlang call streams from the set $\mathbb{C}_{\text{Er},i} = \{1, 2, ..., c_{\text{Er},i}\}$ according to the available set of services. The sources that belong to the set $\mathbb{Z}_{\text{En},j}$ can generate Engset call streams from the set $\mathbb{C}_{\text{En},j} = \{1, 2, ..., c_{\text{En},j}\}$, whereas the sources that belong to the set $\mathbb{Z}_{\text{Pa},k}$ can generate Engset call streams from the set $\mathbb{C}_{\text{Pa},k} = \{1, 2, ..., c_{\text{Pa},k}\}$.

The participation of class c (from the set $\mathbb{M}$) in the structure

ISBN: 978-1-61208-360-5

of traffic generated by sources from the appropriate sets $\mathbb{Z}_{\mathrm{Er},i}$, $\mathbb{Z}_{\mathrm{En},j}$ and $\mathbb{Z}_{\mathrm{Pa},k}$ can be determined by the parameter $\eta_{\mathrm{Er},i,c}$, $\eta_{\mathrm{En},j,c}$ and $\eta_{\mathrm{Pa},k,c}$:

$$\sum_{c=1}^{c_{\mathrm{Er},i}} \eta_{\mathrm{Er},i,c} = 1, \quad \sum_{c=1}^{c_{\mathrm{En},j}} \eta_{\mathrm{En},j,c} = 1, \quad \sum_{c=1}^{c_{\mathrm{Pa},k}} \eta_{\mathrm{Pa},k,c} = 1. \qquad (1)$$

IV. Resource access control mechanisms

A. Bandwidth Reservation

Bandwidth reservation in the switching network consists in introducing the reservation threshold R_c [1][9][12][13]. R_c determines the borderline state of the link (or the group of links), in which servicing the class c calls is still possible. All states higher than R_c belong to the reservation space S_c in which the class c calls are blocked: $S_c = V - R_c$, where V is the total capacity of the group. The reservation mechanism was introduced to the classes that belonged to the set $\mathbb{R}$ which is a sub-set of the set $\mathbb{M}$.

B. Threshold Mechanisms

In switching networks with the applied threshold mechanism the parameters of the offered traffic change depending on the load of the system [6][9][14]. In the algorithm, the threshold mechanism is introduced to outgoing directions only [6]. According to this algorithm, for each class of calls a set of thresholds is individually introduced. For example, for class c we have a set $(Q_{c,1}, Q_{c,2}, \ldots, Q_{c,q})$. In each threshold area u of class c $\{Q_{c,u} < n \leq Q_{c,u+1}\}$ a traffic stream of class c, defined by own set of parameters $\{t_{c,u}, \mu_{c,u}\}$, is offered. Additionally, we assume that $t_{c,0} > t_{c,1} > \ldots > t_{c,u} > \ldots > t_{c,q}$ and $\mu_{c,0}^{-1} \leq \mu_{c,0}^{-1} \leq \ldots \leq \mu_{c,u}^{-1} \leq \ldots \leq \mu_{c,q}^{-1}$.

C. Hysteresis Mechanism

The hysteresis mechanisms also change traffic parameters of carried traffic in relation to the occupancy state of a system [7][9]. In hysteresis algorithm, when the load of the system exceeds the pre-defined limit Q_1, a decrease in the number of assigned BBUs for calls (currently offered and serviced) of classes belonging to the set $\mathbb{H}$ ensues and the average holding time of the call may be increased. When the load of the system is below the hysteresis limit Q_2, this situation is followed by an increase in the number of BBUs allocated to the compressed calls from the set $\mathbb{H}$ and a decrease in the holding time of these calls.

V. General Assumptions

The developed simulator of switching networks was written in C++ using the object programming technique. The process interaction method was used to develop the simulation model. Thus developed simulator allows us to determine values of the blocking probability, loss probability, as well as values of traffic serviced by calls of individual traffic classes, depending on the threshold mechanism involved, in switching networks with point-to-point selection, point-to-group selection and point-to-group selection with a number of attempts to set up a connection. The input data of the simulator are: the capacity and the structure of the switching network. For each traffic class, the number of demanded BBUs, service time and the parameters related to the introduced threshold mechanisms (the demanded number of BBUs in particular threshold areas, threshold boundaries) are given. The sets of traffic sources and their type (Erlang, Engset, Pascal) are defined. In addition, the average value of traffic offered to a single BBU of the system is also given.

In order to perform simulation experiments for a system with the capacity V and composed of switches $\upsilon \times \upsilon$ of links in which the capacity of a single link is f BBU, the values of the following parameters have to be introduced: the number of defined traffic classes m, the number t_c of demanded BBUs necessary to set up a connection of class c and the average service time μ_c^{-1} for a call of class c, the number of sets of traffic sources s_I, s_J, s_K, the sets $\mathbb{Z}_{\mathrm{Er},i}$, $\mathbb{Z}_{\mathrm{En},j}$ and $\mathbb{Z}_{\mathrm{Pa},k}$ of traffic sources, the number of classes $c_{\mathrm{Er},i}$, $c_{\mathrm{En},j}$ and $c_{\mathrm{Pa},k}$ that belong respectively to the sets $\mathbb{C}_{\mathrm{Er},i}$, $\mathbb{C}_{\mathrm{En},j}$ and $\mathbb{C}_{\mathrm{Pa},k}$ of traffic classes, the participation $\eta_{\mathrm{Er},i,c}$, $\eta_{\mathrm{En},j,c}$ and $\eta_{\mathrm{Pa},k,c}$ of calls of class c in traffic generated by sources that belong respectively to the sets $\mathbb{Z}_{\mathrm{Er},i}$, $\mathbb{Z}_{\mathrm{En},j}$ and $\mathbb{Z}_{\mathrm{Pa},k}$ of traffic sources and the number $N_{\mathrm{En},j}$ or $S_{\mathrm{Pa},k}$ of Engset traffic sources from the set $\mathbb{Z}_{\mathrm{En},j}$ and Pascal traffic sources from the set $\mathbb{Z}_{\mathrm{Pa},k}$, threshold boundaries $Q_{c,u}$, defined in the output directions for calls of class c, the number of BBUs $t_{c,u}$ demanded by calls of class c in the threshold area u and the average service time $\mu_{c,u}^{-1}$ for a call of class c in area u.

Additionally, the average traffic a offered to a single BBU in the system is given. On the basis of the above parameters it is possible to determine in the simulator the intensity $\lambda_{\mathrm{Er},i}$, $\gamma_{\mathrm{En},j}$ or $\gamma_{\mathrm{Pa},k}$ of calls generated by the sources of a given type of the traffic stream. In the case of the Engset and Pascal streams, the intensities $\gamma_{\mathrm{En},j}$ and $\gamma_{\mathrm{Pa},k}$ determine the call intensity for calls generated by a single free source. Thus, the parameters $\lambda_{\mathrm{Er},i}$, $\gamma_{\mathrm{En},j}$ and $\gamma_{\mathrm{Pa},k}$ can be determined, depending on the average traffic offered to a single BBU:

$$\lambda_{\mathrm{Er},i} = \frac{aV\upsilon}{S\left[\sum_{c=1}^{c_{\mathrm{Er},i}} t_c \eta_{\mathrm{Er},i,c}\right]\left[\sum_{c=1}^{c_{\mathrm{Er},i}} \mu_c \eta_{\mathrm{Er},i,c}\right]}, \qquad (2)$$

$$\gamma_{\mathrm{En},j} = \frac{aV\upsilon}{S\left[\sum_{c=1}^{c_{\mathrm{En},j}} t_c \eta_{\mathrm{En},j,c}\right]\left[\sum_{c=1}^{c_{\mathrm{En},j}} \mu_c \eta_{\mathrm{En},j,c}\right] N_{\mathrm{En},j}}, \qquad (3)$$

$$\gamma_{\mathrm{Pa},k} = \frac{aV\upsilon}{S\left[\sum_{c=1}^{c_{\mathrm{Pa},k}} t_c \eta_{\mathrm{Pa},k,c}\right]\left[\sum_{c=1}^{c_{\mathrm{Pa},k}} \mu_c \eta_{\mathrm{Pa},k,c}\right] S_{\mathrm{Pa},k}}. \qquad (4)$$

The parameters determined on the basis of (2), (3), and (4) can be treated as the parameters for the exponential distribution that describes the new call arrival process.

With a determination of the blocking probability, the condition for the termination of the simulation experiment is the duration time for individual series necessary to generate a given number of calls of the least active class (most frequently this is a class with the highest number of demanded BBUs). In the case of the loss probability, the condition for the termination of the simulation experiment is the appropriate aggregated number of generated calls of the least active class. The time required to generate a given number of calls is chosen in such a way as to obtain 95% confidence interval. Conventionally, the average result is calculated on the basis of 5 series. In practice, to obtain the confidence interval at the level of 95% it is necessary to generate about 1,000,000 calls of the least active class.

The algorithm according to which the simulation program operates can be written on the basis of the following steps: **1.** Initial configuration of the simulation model – creation of all sources that generate calls of different traffic classes. **2.** Setting the system time to zero. **3.** Activation of traffic sources and events display (call arrival) in the list. **4.** Checking of the simulation termination condition. If the condition is satisfied, then the simulation is terminated and the results are recorded in a file. **5.** Updating of the system time to the time of the appearance of the first event from the list. **6.** Execution of the first event item from the list. **7.** Removal of the first event item from the list and return to step 2.

Two events were defined in the simulation model of the switching network: *the arrival of a new call* and *the termination of call service*. According to the process interaction method, these events are serviced by one function. This function has a different form for each type of Erlang, Engset and Pascal traffic stream. The approach described above makes it possible to define many different traffic classes in the system and to allocate (assign) them to different types of sets of traffic sources. In the initial configuration of the simulation model it is essential to create all sources that generate calls of different traffic classes.

A. Simulation of a system with the sets of Erlang sources

Consider a system in which the set i of Erlang traffic sources $\mathbb{Z}_{\text{Er},i}$ has been defined. In the system, also the set $\mathbb{C}_{\text{Er},i} = \{1, 2, ..., c_{\text{Er},i}\}$ of traffic classes whose calls can be generated by sources from the set $\mathbb{Z}_{\text{Er},i}$ has been defined. In the initial configuration of the system it is necessary to plan ahead the arrival of a call of class c from the set $\mathbb{C}_{\text{Er},i}$. The function that executes events related to the set of Erlang traffic sources can be described in the following way: **1.** Planning of the arrival (appearance) of a new call generated by a source that belongs to the set $\mathbb{Z}_{\text{Er},i}$ according to the exponential distribution where the parameter is the intensity $\lambda_{\text{Er},i}$. Allocation of a given call to class c from the set $\mathbb{C}_{\text{Er},i}$, on the basis of the parameter $\eta_{\text{Er},i,c}$, according to the uniform distribution. Recording the event in the list. **2.** Checking whether the system has sufficient resources to admit a call for service: **a)** Checking whether any of the links of the demanded output direction has at least t_c free BBUs. If not, the call is lost due to the external blocking. **b)** Checking whether there is a path between the input link, at which a call appeared, and the output link of the demanded direction that has at least t_c unoccupied BBUs. If not, the call is lost due to the internal blocking. If any of the conditions (a) or (b) is not satisfied, next steps are omitted. **3.** Occupation of the resources demanded by a call of class c. **4.** Planning of the termination of service according to the exponential distribution where the parameter is the intensity μ_c. Recording the event in the list. **5.** Termination of service and release of resources.

B. Simulation of a system with the sets of Engset sources

Consider a system in which the set j of Engset traffic sources $\mathbb{Z}_{\text{En},j}$ has been defined. In the system, calls from the set $\mathbb{C}_{\text{En},j} = \{1, 2, ..., c_{\text{En},j}\}$ of traffic classes can be generated by $N_{\text{En},j}$ sources from the set $\mathbb{Z}_{\text{En},j}$. It is necessary to plan in the initial configuration of the system an appearance (arrival) of a call of class c from the set $\mathbb{C}_{\text{En},j}$ generated by each of $N_{\text{En},j}$ sources. Hence, the function executing the events related to the set of Engset traffic sources can be presented as follows: **1.** Checking whether the system has sufficient resources to admit a call for service: **a)** Checking whether any of the links of the demanded output link has at least t_c free BBUs. If not, the call is lost due to the external blocking. **b)** Checking whether there is a path between the input link at which a call appears and the output link of the demanded direction that has at least t_c free BBUs. If not, the call is lost due to the internal blocking. If any of the conditions (a) or (b) cannot be satisfied, pass on to step 5. **2.** Occupation of the resources demanded by a call of class c. **3.** Planning of the termination of service according to the exponential distribution where the parameter is the intensity μ_c. Recording the event in the list. **4.** Termination of service and release of resources. **5.** Planning of the appearance (arrival) of a new call generated by a free source from the set $\mathbb{Z}_{\text{En},j}$ of traffic sources according to the exponential distribution where the parameter is the intensity $\gamma_{J,j}$. Allocation of the generated call to class c from the set $\mathbb{C}_{\text{En},j}$ on the basis of the parameter $\eta_{\text{En},j,c}$ according to the uniform distribution. Recording the event in the list.

Unlike the system with the sets of Erlang sources, in the system with the sets of Engset sources a generation of a new call is possible only when there is a free source (from among $N_{\text{En},j}$ sources) capable of generating this call. This condition will be satisfied in a situation when the call service process for another call is completed. Hence, in the function related to event service, a termination of service for a call is immediately followed by a possibility to plan an arrival of another call.

C. Simulation of a system with the sets of Pascal sources

Consider a system in which a set of Pascal traffic sources $\mathbb{Z}_{\text{Pa},k}$ has been defined. The system also has a defined set $\mathbb{C}_{\text{Pa},k} = \{1, 2, ..., c_{\text{Pa},k}\}$ of traffic sources whose calls can be generated by $S_{\text{Pa},k}$ sources from the set $\mathbb{Z}_{\text{Pa},k}$. In the initial configuration of the system it is necessary to plan the appearance (arrival) of a call of class c from the set $\mathbb{C}_{\text{Pa},k}$, generated by each of $S_{\text{Pa},k}$ sources. Therefore, the function executing the events related to the set of Pascal traffic sources leads to the execution of the following task: **1.** Planning of the arrival of a new call, generated by a source from the set $\mathbb{Z}_{\text{Pa},k}$, according to the exponential distribution where the parameter is the intensity $\gamma_{\text{Pa},k}$. Allocation of the call to class c from the set $\mathbb{C}_{\text{Pa},k}$, on the basis of the parameter $\eta_{\text{Pa},k,c}$, according to the uniform distribution. Recording the event in the list. **2.** Checking whether the system has sufficient resources to service the call: **a)** Checking whether any of the links of the demanded output direction has at least t_c free BBUs. If not, the call is lost due to the external blocking. **b)** Checking whether there is a path between the input link at which a call appears (arrives) and the output link of the demanded output direction that has at least t_c free BBUs. If not, the call is lost due to the internal blocking. If any of the conditions (a) or (b) is not satisfied, the next steps are omitted. **3.** Occupation of the resources that are demanded by a call of class c. **4.** Planning of the termination of service according to the exponential distribution where the parameter is the intensity μ_c. Recording the event in the list. **5.** Addition of two sources (at the moment of the admittance of a given call). Planning of the arrival of new calls generated by new sources from the set $\mathbb{Z}_{\text{Pa},k}$ according to the exponential distribution where the parameter

ISBN: 978-1-61208-360-5

is the intensity $\gamma_{\mathrm{Pa},k}$. Allocation of new calls to class c from the set $\mathbb{C}_{\mathrm{Pa},k}$ on the basis of the parameter $\eta_{\mathrm{Pa},k,c}$ according to the uniform distribution. Recording the event in the list. **6.** Termination of service and release of resources. **a)** Removal of two sources, currently not serviced, at the moment of a termination of service of a given call. **b)** Removal of the events related to the removed source.

VI. Simulation experiments of switching networks with threshold mechanisms

The simulator makes it possible to examine the influence of different factors on the values of the blocking probability and on the values of carried traffic. The factors involved include: the number of attempts to set up a connection and the link occupation strategy. This section presents an extensive set of results that determine the influence of different parameters on the characteristics of switching networks such as, for example, the applied threshold mechanism, the number of attempts to set up a connection, and the link occupation strategy. The results of the simulation experiments are presented in the from of graphs with confidence intervals that have been determined on the basis of the t-Student distribution (with 95-percent confidence interval) for 5 series. The duration time for each of the series has been determined on the basis of the time required to generate 1,000,000 calls of the least active class. In each case, the confidence interval does not exceed 5% of the average value of the result of the simulation experiment.

A. *Influence of the applied threshold mechanism*

To examine the influence of the applied resource access control mechanism on the values of carried traffic and the blocking probability simulation experiments were performed. Three types of mechanisms were considered: reservation, threshold and hysteresis mechanisms.

The study was carried out for the following parameters: **structure of switching network**: $v = 4$, $f = 32$ PJP, $V = 128$ PJP; **structure of offered traffic**: traffic classes: $m = 3$, $t_{1,0} = 1$ PJP, $\mu_{1,0}^{-1} = 1$, $t_{2,0} = 4$ PJP, $\mu_{2,0}^{-1} = 1$, $t_{3,0} = 8$ PJP, $\mu_{3,0}^{-1} = 1$; sets of traffic sources: $S = 3$, $\mathbb{C}_{\mathrm{Er},1} = \{1\}$, $\eta_{\mathrm{Er},1} = 1$, $\mathbb{C}_{\mathrm{En},2} = \{2\}$, $\eta_{\mathrm{En},2} = 1$, $N_{\mathrm{En},2} = 128$, $\mathbb{C}_{\mathrm{Pa},3} = \{3\}$, $\eta_{\mathrm{Pa},3} = 1$, $S_{\mathrm{Pa},3} = 128$; **reservation mechanism**: $R_1 = R_2 = 90$ PJP, $\mathbb{R} = \{1, 2\}$; **threshold mechanism**: $t_{2,1} = 2$ PJP, $\mu_{2,1}^{-1} = 2$, $t_{3,1} = 4$ PJP, $\mu_{3,1}^{-1} = 2$, $q_3 = 1$, $Q_{3,1} = 90$ PJP; **hysteresis mechanism**: $t_{2,1} = 2$ PJP, $\mu_{2,1}^{-1} = 2$, $t_{3,1} = 4$ PJP, $\mu_{3,1}^{-1} = 2$, $Q_1 = 100$ PJP, $Q_2 = 80$ PJP, $\mathbb{H} = \{2, 3\}$;

In the case when the hysteresis mechanism or the threshold mechanism have been applied to the switching networks under consideration we can observe the highest values of carried traffic as compared to a switching network without those mechanisms being introduced. A reverse situation is to be found in the case of switching networks with reservation mechanisms (Figure 1).

In switching networks in which a proper threshold mechanism or a threshold mechanism with hysteresis have been introduced it was observable that in the case of classes that did not undergo these mechanisms the blocking probability for low intensities of traffic was lower than the blocking probability in networks without the introduced mechanisms. With the traffic intensity being increased, a reverse situation was observed (Figure 2). For classes that undergo the introduced proper threshold mechanisms or threshold mechanisms with hysteresis, it was observed that the blocking probability was always lower than the blocking probability in networks without introduced relevant mechanisms (Figure 3). This behavior of the system results from the fact that calls of classes that undergo the threshold mechanisms are allocated the number of BBUs that is not higher, and in many cases lower, than in the case of switching networks without threshold mechanisms.

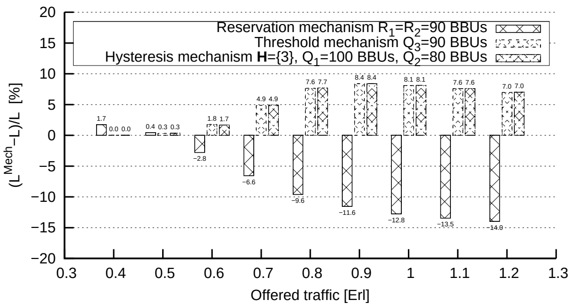


Figure 1. Percentage change in the value of carried traffic in relation to the applied resource access control mechanism

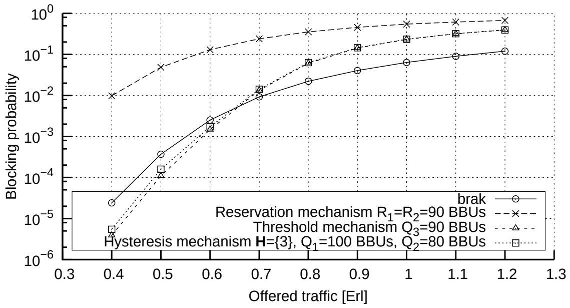


Figure 2. Blocking probability for calls of class 2 in relation to the applied resource access control mechanism

The blocking probability in switching networks with reservation mechanisms for traffic classes that undergo these mechanisms is always higher than the blocking probability in networks with no introduced mechanisms (Figure 2). Whereas in the case of classes that do not undergo reservation mechanisms (privileged classes), the blocking probability is always lower than the blocking probability in networks without new call admission control mechanisms (Figure 3). Such a behavior of

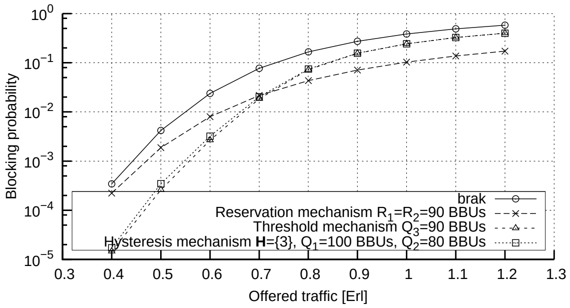


Figure 3. Blocking probability for calls of class 3 in relation to the applied resource access control mechanism

ISBN: 978-1-61208-360-5

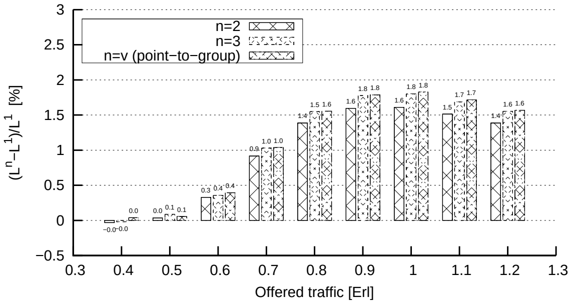


Figure 4. Percentage change in the value of carried traffic in relation to the number of attempts to set up a connection

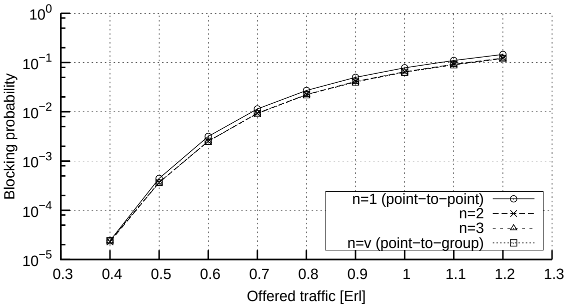


Figure 5. The blocking probability for calls of class 2 in relation to the number of attempts to set up a connection

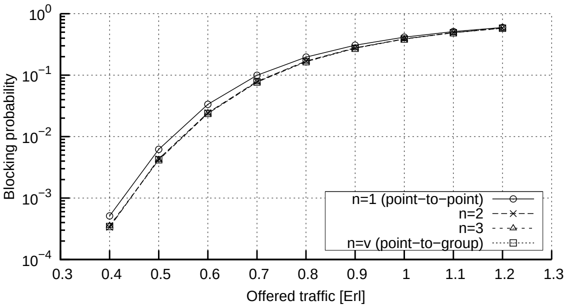


Figure 6. The blocking probability for calls of class 3 in relation to the number of attempts to set up a connection

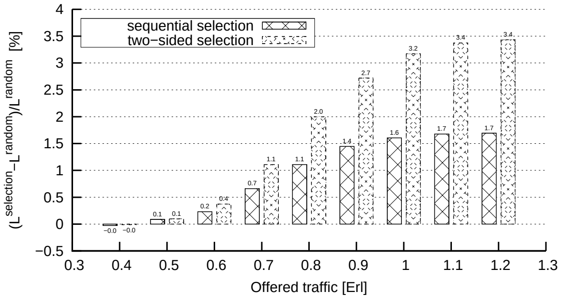


Figure 7. Percentage change in the value of carried traffic in relation to the link occupation strategy

the system results from the fact that for classes for which reservation boundaries have been defined the area of available network resources has been thus limited (decreased). In turn, for classes that do not undergo the reservation mechanism the whole capacity of the system is available.

B. Influence of the number of attempts

As a result of the simulation experiments carried out in the course of the study it was possible to examine the influence of the number of attempts to set up a connection inside the switching network on the values of carried traffic and the blocking probability. The maximum number of attempts to set up a connection in the network is equal to the number of links of a given output direction and is υ. The number of attempts is equal to 1 and corresponds to the point-to-point selection in the switching network, while υ attempts correspond to the point-to-group selection.

The study was performed for the following structure of the switching network: structure of switching network: $\upsilon = 4$, $f = 32$ PJP, $V = 128$ PJP; traffic classes: $m = 3$, $t_{1,0} = 1$ PJP, $\mu_{1,0}^{-1} = 1$, $t_{2,0} = 4$ PJP, $\mu_{2,0}^{-1} = 1$, $t_{3,0} = 8$ PJP, $\mu_{3,0}^{-1} = 1$; sets of traffic sources: $S = 3$, $\mathbb{C}_{\mathrm{Er},1} = \{1\}$, $\eta_{\mathrm{Er},1} = 1$, $\mathbb{C}_{\mathrm{En},2} = \{2\}$, $\eta_{\mathrm{En},2} = 1$, $N_{\mathrm{En},2} = 128$, $\mathbb{C}_{\mathrm{Pa},3} = \{3\}$, $\eta_{\mathrm{Pa},3} = 1$, $S_{\mathrm{Pa},3} = 128$;

While analyzing the systems under consideration in view of the influence of the number of attempts to set up a connection in the network on the value of carried traffic and the values of blocking probabilities, a number of interesting observations can be drawn up.

In many cases, only one attempt to set up a connection between a given input and a given output is not sufficient to determine a free connection path. The probability of finding a free connecting path increases with the increase in the number of attempts of setting up a connection. The highest values of carried traffic have been then observed for networks in which the number of attempts to set up a connection is υ (Figure 4). The blocking probability is the lowest in networks in which the number of attempts to set up a connection is υ. The biggest differences in the values of the blocking probabilities were observed between 1 and 2 attempts (Figures 5-6).

C. Influence of the link occupation strategy

The last factor influencing the changes in the values of carried traffic and the values of blocking probabilities to be examined in the study was the link occupation strategy. The relevant simulation experiments were performed to evaluate the influence of the adopted link occupation strategy on the values of carried traffic and the blocking probability. The experiments involved the following strategies: random selection (inter-stage links and output links of the switching network are randomly selected), sequential selection (the selection of inter-stage links of the network and output links begins with the first free link, while links are numbered from 1 to υ), and the so-called two-sided selection (calls of classes with lower demands begin the occupation of links from the numbers $1, 2, ...$, whereas calls of classes that demand the highest number of BBUs to set up a connection occupy links starting from the numbers $\upsilon, \upsilon-1, ...$).

The study was performed for the following structure of the switching network: structure of switching network: $\upsilon = 4$, $f = 32$ PJP, $V = 128$ PJP; traffic classes: $m = 3$, $t_{1,0} = 1$ PJP, $\mu_{1,0}^{-1} = 1$, $t_{2,0} = 4$ PJP, $\mu_{2,0}^{-1} = 1$, $t_{3,0} = 8$ PJP, $\mu_{3,0}^{-1} = 1$; sets of traffic sources: $S = 3$, $\mathbb{C}_{\mathrm{Er},1} = \{1\}$, $\eta_{\mathrm{Er},1,1} = 1$, $\mathbb{C}_{\mathrm{En},2} = \{2\}$, $\eta_{\mathrm{En},2,2} = 1$, $N_{\mathrm{En},2} = 128$, $\mathbb{C}_{\mathrm{Pa},3} = \{3\}$, $\eta_{\mathrm{Pa},3,3} = 1$, $S_{\mathrm{Pa},3} = 128$;

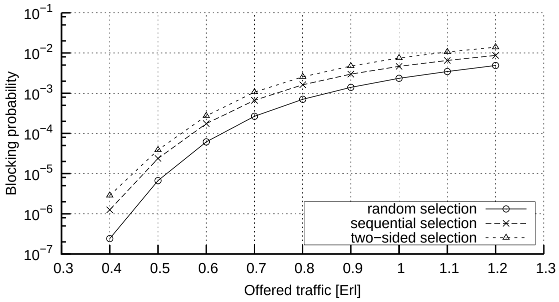


Figure 8. Blocking probability for calls of class 1 in relation to the link occupation strategy

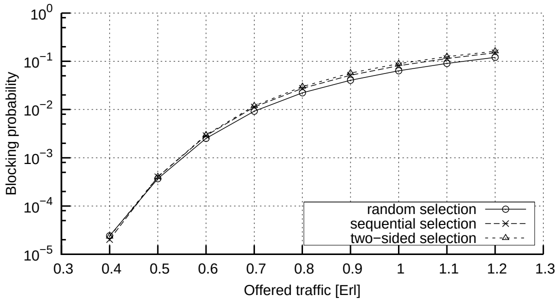


Figure 9. Blocking probability for calls of class 2 in relation to the link occupation strategy

The highest values of carried traffic in the switching network with point-to-group selection were observed for networks with the two-sided selection, whereas the lowest values for networks with the random selection of links (Figure 7). In the analysis of the influence of the link occupation strategy on the blocking probability, for the classes that demanded the highest number of BBUs, the lowest values were observed in the networks with two-sided selection (Figure 10). A reverse situation was recorded for the remaining traffic classes (Figures 8, 9).

VII. Conclusion

The papers presents the structure of the simulator of multi-service switching networks with multi-service sources. The developed simulator allows us to determine an influence of the number of attempts to set up a connection and the link occupation strategy on the values of the blocking probability and on the values of carried traffic. In the simulator, a selection of various resource management algorithms has been implemented. The obtained results makes it possible to select the right resource management mechanism to realize an assumed policy (e.g., equalisation of blocking probability of different traffic classes, maximization of the value of carried traffic, etc.). The simulator is also an indispensable tool for evaluation of analytical models of multi-service switching networks with multi-service sources.

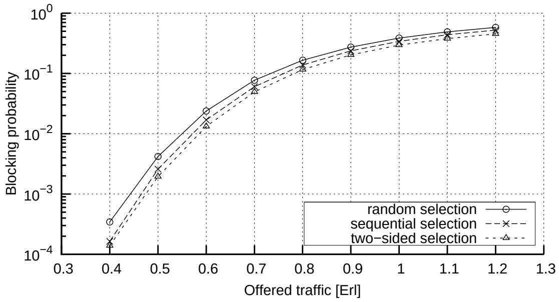


Figure 10. Blocking probability for calls of class 3 in relation to the link occupation strategy

References

[1] I. D. Moscholios and M. D. Logothetis, "The erlang multirate loss model with batched poisson arrival processes under the bandwidth reservation policy," Computer Communications, vol. 33, no. 1, Nov. 2010, pp. S167–S179.

[2] J. S. Vardakas, I. D. Moscholios, M. D. Logothetis, and V. G. Stylianakis, "An analytical approach for dynamic wavelength allocation in wdm-tdma pons servicing on-off traffic," IEEE/OSA Journal of Optical Communications and Networking, vol. 3, no. 4, Apr. 2011, pp. 347–358.

[3] I. D. Moscholios, V. G. Vassilakis, J. S. Vardakas, and M. D. Logothetis, "Retry loss models supporting elastic traffic," Advances in Electronics and Telecommunications, vol. 2, no. 3, Oct. 2011, pp. 8–13.

[4] M. Głąbowski, M. Sobieraj, and P. Zwierzykowski, "Modeling of Resource Managament Mechanism for Virtual Networks," in Information Systems Architecture and Technology, service oriented networked systems ed., A. Grzech, L. Borzemski, J. Świątek, and Z. Wilimowska, Eds. Oficyna Wydawnicza Politechniki Wrocławskiej, 2011, pp. 303–316.

[5] M. Głąbowski, "Continuous threshold model for multi-service wireless systems with PCT1 and PCT2 traffic," in Proceedings of 7th International Symposium on Communications and Information Technologies, Sydney, 2007, pp. 427–432.

[6] M. Głąbowski and M. Sobieraj, "Point-to-group blocking probability in switching networks with threshold mechanisms," in roceedings of the Fifth Advanced International Conference on Telecommunications. Venezia: IEEE Computer Society, 2009, pp. 95–100.

[7] M. Głąbowski, M. Sobieraj, M. Stasiak, and J. Weissenberg, "Switching networks with hysteresis mechanism," in Proceedings of the The Seventh Advanced International Conference on Telecommunications, M. Głąbowski and D. Mynbaev, Eds., St. Maarten, The Netherlands Antilles, 2011, pp. 135–140.

[8] M. Głąbowski and M. Sobieraj, "Multi-service Switching Networks with Resource Management Mechanisms," The Mediterranean Journal of Computers and Networks, vol. 7, no. 4, 2011, pp. 292–303.

[9] ——, "Call admission control mechanisms in multi-service switching networks with bpp traffic," in Proceedings of III International Interdisciplinary Technical Conference of Young Scientists, Poznan, Poland, 2010, pp. 240–244.

[10] M. Głąbowski and A. Kaliszan, "Simulator of full-availability group with bandwidth reservation and multi-rate bernoulli-poisson-pascal traffic streams," in Proceedings of Eurocon 2007, Warsaw, 2007, pp. 2271–2277.

[11] M. Głąbowski, M. Sobieraj, and M. Stasiak, "Modeling switching networks with multi-service sources and point-to-group selection," in 18th Asia-Pacific Conference on Communications (APCC 2012), Jeju Island, Korea, 2012, pp. 686–691.

[12] ——, "Modelling limited-availability groups with bpp traffic and bandwidth reservation," in Proceedings of the Fifth Advanced International Conference on Telecommunications. Venezia: IEEE Computer Society, 2009, pp. 89–94.

[13] I. D. Moscholios, J. S. Vardakas, M. D. Logothetis, and M. N. Koukias, "A quasi-random multirate loss model supporting elastic and adaptive traffic under the bandwidth reservation policy," Int. Journal on Advances in Networks and Services, vol. 6, no. 3 and 4, Dec. 2013, pp. 163–174.

[14] V. G. Vassilakis, I. D. M. G. A. Kallos, and M. D. Logothetis, "On call admission control in w-cdma networks supporting handoff traffic," Ubiquitous Computing and Communication Journal, Sep. 2009, pp. S167–S179.

Trust-based Incentive Cooperative Relay Routing Algorithm for Wireless Networks

Youngjae Park
Department of Computer Science
Sogang University
Seoul, Korea
yjpark903@sogang.ac.kr

Sungwook Kim
Department of Computer Science
Sogang University
Seoul, Korea
swkim01@sogang.ac.kr

***Abstract*—Recently, cooperative communication has been proposed as an effective approach to enhance system performance in wireless networks. Cooperative communications fundamentally change the abstraction of a wireless link and offer significant potential advantages for wireless communication networks. In this paper, we propose a novel cooperative relay routing scheme based on the trust-based incentive mechanism. To maximize the network performance, the proposed scheme can take into account the measure of the probability of a relay node succeeding at a given relay service. By considering the current network condition, we can select the most adaptable relay node and pay the incentive-price for relay service. Evidences from simulations demonstrate that the proposed scheme outperforms the existing schemes.**

Keywords-Cooperative communication; Relay selection; Trust evaluation; Incentive mechanism; Vickrey-Clarke-Groves (VCG) mechanism.

I. INTRODUCTION

Today, wireless communication systems are primarily designed based on point-to-point links whose performance is limited by the resources of a single transmitter. In particular, bandwidth and transmit power constraints often prevent a source node from achieving a desired data rate or communication range. Cooperative communications have emerged as a new novel approach that enables a source to tap into the available resources of local neighboring nodes in order to increase throughput, range and covertness [1]. Therefore, a cooperative communication improves performance in wireless systems, but it requires some nodes to expend energy acting as relays. Since energy is scarce, wireless nodes refuse to cooperate in order to conserve resources. However, only when relay node expends extra energy on its behalf, network performance can increase. Due to this realistic constraint, cooperative communication has yet to see widespread use in practical wireless systems [2]. Therefore, it is necessary to design incentive-aware relay routing algorithms for stimulating cooperation among nodes. But, despite the concerns, not much work has been done in this direction. In this article, we address this problem by examining the trust based incentive mechanism. In the proposed scheme, the attenuation window technique is adopted to estimate the trust value based on historical data [3]. Based on the individual trust value, we select an adaptive relay node and cooperative routing takes place with the help of the relay node, which is incentivized by the relay service price. To calculate the service price for a relay node, the basic concept of *Vickrey-Clarke-Groves* (*VCG*) mechanism [4] is used.

Usually, mechanisms that implement efficient social choice functions in environments in which participants have private information about their preferences have been studied extensively in the economics literature. A well-known class of such mechanisms are the *VCG* mechanisms [5]. It is a generalization of the famous *Vickery* auction where bidders submit written bids without knowing the bid of the other people in the auction. The important property of the *VCG* mechanism is that it is truthful; each bidder reveals his/her true value no matter what strategies the other bidder chooses. The main result in this setting provides an incentive compatible for the most natural social choice function [5].

In this work, the proposed scheme adapts in order to obtain the trust values in the real-time, online. This approach is able to select the relay node under dynamic changing wireless network environments; it is essential in order to maximize the network performance. Recently, several cooperative communication schemes – the *Threshold based Cooperative Communication* (*TCC*) scheme [6] and the *Distributed Relay Routing* (*DRR*) scheme [7] - have been presented for wireless network systems. This research implicates that cooperative communication is envisioned to achieve high diversity gain in terms of outage probability and outage capacity and scaling network capacity. The *TCC* scheme [6] is a new approach to relay selection using a threshold-based transmission protocol for a wireless system. The *DRR* scheme [7] is a decentralized and localized algorithm for joint dynamic routing, relay assignment, and spectrum allocation in a distributed and dynamic environment. All the earlier work has attracted a lot of attention and introduced unique challenges to efficiently solve the cooperative communication problem. Compared to these schemes [6][7], the proposed scheme attains better performance in cooperative communications.

This paper is organized as follows. Section II presents the proposed algorithms in detail. In Section III, performance evaluation results are presented along with comparisons with

ISBN: 978-1-61208-360-5

other schemes. Finally, concluding remarks are given in Section IV.

II. Proposed Cooperative Communication Algorithm

In this paper, trust-based *VCG* mechanism is used to design a new cooperative communication algorithm. Based on the trust value of each relay node, the most adaptable relay node is selected. To induce selfish relay nodes to participate cooperative communications, the relay service price is provided for the relay service. Finally, an effective solution can be obtained in the constantly changing environment.

A. *Trust Evaluation for Relay nodes*

In this paper, we assume a wireless network situation where there is a set (*N*) of potential relay nodes, $N = \{1,2..,i..,n\}$. Each relay node (i.e., $i \in N$) has a service value and privately-known relay cost of performing the service request τ. Let σ denote a particular relay service acceptance within the space of possible acceptances Ψ and τ^i represent that the node *i* allows to relay call service τ; $\sigma \in \Psi$ and $\Psi = \{\emptyset, \tau^1, \tau^2 ... \tau^n\}$, where $\emptyset$ denotes the case where the relay request is rejected. If $\Psi = \{\tau^i\}$, the relay node *i* participates the cooperative communication. The set of available power levels ($\mathbb{S}$) for each relay node is assumed as below.

$$\mathbb{S} = \{\Pi_{i \in N} p_i | p_i \in [p_{min}, p_{max}]\} \tag{1}$$

where p_i is the power level of relay node *i*. The p_{min}, p_{max} are the pre-defined minimum and maximum power levels, respectively. Each relay node selects a power level from the $\mathbb{S}$, and estimates the expected value ($v_i(\tau)$) as follows.

$$v_i(\tau) = \left(W \times \log_2 \left(1 + \frac{\gamma_i(\mathbb{P})}{\Omega} \right) \right) / p_i$$
$$\text{s.t., } \gamma_i(\mathbb{P}) = \frac{p_i h_{ii}}{\vartheta_i + \sum_{j \neq i} p_j h_{ji}} \tag{2}$$

where $\mathbb{P}$ is the power level vector for all nodes and *W* is the channel bandwidth of relay node *i*, and Ω (Ω ≥ 1) is the gap between uncoded M-ary Quadrature Amplitude Modulation (M-QAM) and the capacity, minus the coding gain [6]. Usually, service value is defined as the number of information bits that are transmitted without error per unit-time. In wireless networks, it can be achieved with the Signal to Interference plus Noise Ratio (SINR) in the effective range. Therefore, to estimate the service value, the SINR should be obtained. The $\gamma_i(\mathbb{P})$ is a general formula for the relay *i*'s SINR, where ϑ_i is the background noise within the relay node *i*'s bandwidth, h_{ji} is the path gain from the node *j* to the node *i* [8].

Under a dynamically changing network environment, there exists uncertainty about relay nodes successfully completing their assigned relay services. In the proposed scheme, we take into account the trust value (*T*) of relay nodes. $T_i(t)$ is the relay node *i*'s trust value at the time *t*. After the t^{th} iteration, $T_i(t)$ is using the number of packets successfully serviced in the relay node *i* (α_t^i) divided by the total number of packets that have been sent from the source node to the relay node *i* ($\alpha_t^i + \beta_t^i$).

$$T_i(t) = \frac{\alpha_t^i}{\alpha_t^i + \beta_t^i} \tag{3}$$

$T_i(t)$ is a general average function over the whole span of communication historical records. However, for a long-term period evaluation, the α_t^i and β_t^i will be accumulated and are growing into a very large value. In such case, a small amount of the recent malicious behaviours will be hard to be counted and thus has impact on the overall rating of trust. To solve this problem, attenuation window was introduced [3]. By considering more on the up-to-date records, we can calculate the trust value ($T_i(t)$) while fade away the out-of-date records. Based on the attenuation window, the α_t^i and β_t^i values is obtained as below.

$$\alpha_t^i = \sum_{\lambda=k}^{n} e^{-\left(\frac{n+m-t(\lambda)}{c}\right)}$$
$$and \;\; \beta_t^i = \sum_{\lambda=j}^{m} e^{-\left(\frac{n+m-t(\lambda)}{c}\right)} \tag{4}$$

where the *e* is Euler's constant, and *c* is the coefficient to adjust the speed of decreasing in the results of α_t^i and β_t^i. The *n* and *m* are the total number of successfully serviced and non-successfully serviced packets, respectively. The *k* and *j* are the most out-of-date time for successfully serviced and non-successfully serviced packets, respectively. $t(\lambda)$ is the time *t* when λ occurs. For example, there are 3 successful service records regarding to the packet relaying but 2 non-successful service records, i.e., *n*=3 and *m*=2. Here, the successful service time set are *t* = {1, 3, 5} and non-successful service time set are *t* = {2, 4}. Thus the *k*=1 and *j*=2. As the time *t* is from ascending order that it reflects from oldest to latest in time sequence [2]. While *t* is growing bigger and bigger, the value of (*n*+*m*-*t*) will become smaller and smaller, and finally $e^{-\left(\frac{n+m-t}{c}\right)}$ has a strong impact on the recent information. Moreover, the bigger value of coefficient *c*, the slower in speed of decreasing slopes of the value in $e^{-\left(\frac{n+m-t}{c}\right)}$ between 0 and 1. In such way, attenuation window can emphasize the most up-to-date records and fade away the out-of-date records by the speed controlled by the coefficient *c*.

B. *Relay Selection and Relay Service Price Computation*

The proposed scheme adopts the basic concept of *T-VCG* mechanism to provide a normative guide for the payments of relay service [4]. Even more importantly, we consider the trust value in computing the relay payment. With the estimated trust value, the expected payoff value of relay node, $\bar{\chi}(\tau,\sigma,\mathbb{T})$, can be calculated as:

$$\bar{\chi}(\tau,\sigma,\mathbb{T}) = v_\sigma(\tau) \times T_\sigma(\tau),\ \text{s.t.}, \sigma \in \Psi \qquad (5)$$

where $v_\sigma(\tau)$ is the expected value with relay task τ and $T_\sigma(\tau)$ is the trust value in the selected relay node σ, which performs the requested relay service τ. To implement the execution uncertainty by a given relay node, the network system needs to require relay nodes to report their trust value. $\mathbb{T} = \langle T_1(\tau) \ldots T_\sigma(\tau) \ldots T_n(\tau)\rangle$ is the vector of trust values of all the relay nodes. In this paper, $\widehat{\mathbb{T}}$ represents the vector of reported trust values $\langle \widehat{\mathbb{T}}_1(\tau),\ldots,\widehat{\mathbb{T}}_n(\tau)\rangle$; the superscripting the latter with '^' indicates that nodes can misreport their true types. With the expected payoff value, service execution cost is necessary to estimate the total profit. The cost function defines the instantaneous expense for the relaying service σ. It would be a linear function of the tower level, and given by $K \times (p_\sigma)^q$, where K and q are estimation parameters and p_σ is the power level for the σ service.

In the proposed scheme, relay connections are adaptively controlled based on the accurate analysis of costs and payoffs to select the most suitable relay node. In more detail, the relay selection is determined as follows to maximize system efficiency.

$$S^*\big(\Psi,\widehat{\mathbb{T}}\big) = arg \max_{\sigma\in\Psi} \left[\bar{\chi}(\tau,\sigma,\widehat{\mathbb{T}}) - K \times (\hat{p}_\sigma)^q\right] \qquad (6)$$

In the real world operation, $\mathbb{T}$ and p_σ can be misreported (i.e., $\widehat{\mathbb{T}}$ and $\hat{p}_\sigma$). After selecting a relay node, the next step is to compute the relay price, which is an incentive to encourage relay communications. In this paper, the relay service price is similar to that of the traditional *VCG* mechanism in that the marginal contribution of the selected relay node to the wireless network system; it is extracted by comparing the second best decision, excluding the selected relay node. Without the best relay node $S^*\big(\Psi,\widehat{\mathbb{T}}\big)$, the second-best expected payoff for the relay service ($EU_{\sigma\in\Psi_{-S^*(\Psi,\widehat{\mathbb{T}})}}\big(\Psi,\widehat{\mathbb{T}}\big)$) is given by

$$EU_{\sigma\in\Psi_{-S^*(\Psi,\widehat{\mathbb{T}})}}\big(\Psi,\widehat{\mathbb{T}}\big) = \max_{\sigma\in\Psi_{-S^*(\Psi,\widehat{\mathbb{T}})}} \big(\bar{\chi}(\tau,\sigma,\widehat{\mathbb{T}}) - K \times (\hat{p}_\sigma)^q\big) \qquad (7)$$

where $\Psi_{-S^*(\Psi,\widehat{\mathbb{T}})}$ is the set of possible acceptances (Ψ) excluding the best relay node $S^*\big(\Psi,\widehat{\mathbb{T}}\big)$. If the selected relay node ($S^*\big(\Psi,\widehat{\mathbb{T}}\big)$) can success to provide relay service, the relay service price ($RSP_{S^*(\Psi,\widehat{\mathbb{T}})}$) is achieved based on the expected marginal contribution, which is the difference between the best and the second-best expected payoff.

$$RSP_{S^*(\Psi,\widehat{\mathbb{T}})} = \left[\bar{\chi}\big(\tau,S^*\big(\Psi,\widehat{\mathbb{T}}\big),\widehat{\mathbb{T}}\big) - K \times (\hat{p}_{\,S^*(\Psi,\widehat{\mathbb{T}})})^q\right] - EU_{\sigma\in\Psi_{-S^*(\Psi,\widehat{\mathbb{T}})}}\big(\Psi,\widehat{\mathbb{T}}\big) \qquad (8)$$

Sometimes, the selected relay node (i.e., $S^*\big(\Psi,\widehat{\mathbb{T}}\big)$) can fail to provide relay service. In this case, the $RSP_{S^*(\Psi,\widehat{\mathbb{T}})}$ is given by.

$$RSP_{S^*(\Psi,\widehat{\mathbb{T}})} = -\,EU_{\sigma\in\Psi_{-S^*(\Psi,\widehat{\mathbb{T}})}}\big(\Psi,\widehat{\mathbb{T}}\big) \qquad (9)$$

Finally, $RP_i\big(\widehat{\mathbb{C}},\widehat{\mathbb{P}},\sigma\big)$ can be obtained as follows by considering success and fail cases .

$$\begin{aligned} RSP_{S^*(\Psi,\widehat{\mathbb{T}})} &= T_{S^*(\Psi,\widehat{\mathbb{T}})}(t) \\ &\times (\left[\bar{\chi}\big(\tau,S^*\big(\Psi,\widehat{\mathbb{T}}\big),\widehat{\mathbb{T}}\big) - K \times \left(\hat{p}_{\,S^*(\Psi,\widehat{\mathbb{T}})}\right)^q\right] \\ &-EU_{\sigma\in\Psi_{-S^*(\Psi,\widehat{\mathbb{T}})}}\big(\Psi,\widehat{\mathbb{T}}\big)) \\ &+ \left(1 - T_{S^*(\Psi,\widehat{\mathbb{T}})}(t)\right) \times \left[-\,EU_{\sigma\in\Psi_{-S^*(\Psi,\widehat{\mathbb{T}})}}\big(\Psi,\widehat{\mathbb{T}}\big)\ \right] \end{aligned} \qquad (10)$$

III. Performance Evaluation

In this section, the effectiveness of the proposed scheme is validated through simulation. Recently, the TCC scheme [6] and the DRR scheme [5] have been published and introduced unique challenges for cooperative communications. Using a simulation model, we compare the performance of the proposed scheme with these existing schemes to confirm the superiority of the proposed approach. The assumptions implemented in simulation model are as follows. Each relay service is considered Constant Bit Rate (CBR) traffic with having a different deadline. In order to adaptively adjust the control parameters, we partition the time-axis into equal intervals of length (i.e., a short time duration). Every interval, the current system condition is examined periodically by a real-time online approach, and the performance measures obtained on the basis of 50 simulation runs. Relay transmitters use variable-rate M-QAM, with a bounded probability of symbol error and trellis coding with a nominal coding gain. Table 1 shows the system parameters used in the simulation.

For dynamically changing network environments, the design goal of the proposed scheme is to improve the overall network performance. Figures 1 and 2 show the performance comparison for the network throughput and energy efficiency, respectively. All the schemes have similar trends. However, the proposed scheme constantly monitors the current network condition and effectively operates relay routing through the trust based incentive mechanism. Under various network operation times, the system performance of the proposed scheme is better than the other schemes.

TABLE I. SYSTEM PARAMETERS USED IN THE SIMULATION EXPERIMENTS.

Parameter	Value	Description	
N	5	number of relay stations (n)	
W	256 Kbps	bandwidth requirement for service	
P_{min}, P_{max}	50mW, 100mW	pre-defined minimum and maximum power levels	
c	1	control the decreasing slopes of e curve	
K	1	power parameter power cost function	
q	0.7	cost function parameter about power	
ϑ	1×10^{-10}	AWGN background noise	
Ω	1	gap between uncoded M-QAM and the capacity, minus the coding gain	
Parameter	**Initial**	**Description**	**Values**
P_i	50mW	communication power for the users i	50,60,70,80,90,*100mW*

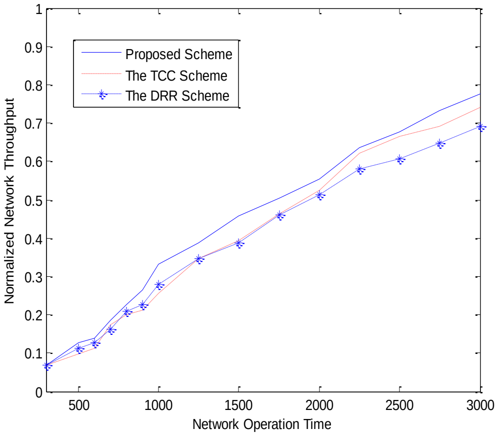


Figure 1. Network Throughput

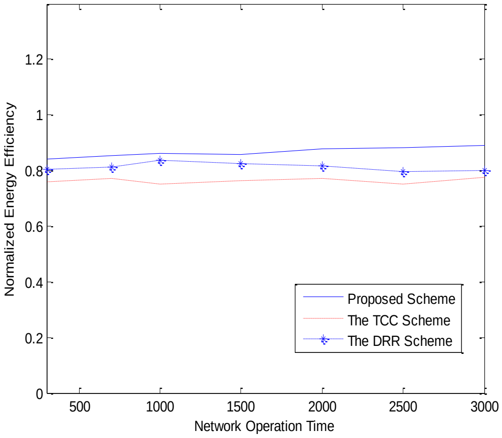


Figure 2. Energy Efficiency

IV. CONCLUSIONS AND FUTURE WORK

With the rapid growth of mobile Internet, offering seamless connectivity and high-speed multimedia services in different types of wireless networks are important features in next generation wireless networks (4G networks). For next-generation wireless networks, cooperative communication is an emerging technology to overcome the current limitations of traditional wireless systems. This promising technique has been considered in the IEEE 802.16 standard, and is expected to be integrated into LTE multi-hop cellular network. In this paper, we have introduced a new trust-based incentive relay routing algorithm for wireless networks. Based on the trust-based VCG mechanism, the proposed algorithm dynamically estimates relay nodes' trust levels and adaptively selects the most adaptable relay node for the data transmission. Moreover, we suggest the new trust estimation mechanism. Our model is designed to adapt to various changes, such as changes in trust behaviors and trust accuracy requirement. Usually, traditional routing methods cannot solve the problem of malicious behavior and build the trusted transfers route between nodes. In the proposed scheme, the subjective confidence for the routing behavior has transform into trust evaluation with the probability model to solve the trust measurement and routing. Simulation results clearly indicate that the proposed algorithm generally exhibits superior performance compared with the other existing schemes. Future work in progress is to study efficient power control and relay selection schemes for green cooperative communications based on the game theory.

REFERENCES

[1] M. Nokleby, and B. Aazhang, "User Cooperation for Energy-Efficient Cellular Communications", IEEE ICC'2010, 2010, pp. 1-5.

[2] M. Xiang, "Trust-based energy aware geographical routing for smart grid communications networks", Master's Thesis, Auckland University of Technology, 2013

[3] S. Kim, "Adaptive Call Admission Control Scheme for Heterogeneous Overlay Networks", Journal of Communications and Networks,vol. 14, no. 4, 2012, pp. 461 – 466.

[4] L. Viet-Anh, R. A. Pitaval, S. Blostein, T. Riihonen, and R. Wichman, "Green cooperative communication using threshold-based relay selection protocols", ICGCS'2010, 2010. pp. 521-526.

[5] D. Lei, T. Melodia, S. N. Batalama, and J. D. Matyjas, "Distributed Routing, Relay Selection, and Spectrum Allocation in Cognitive and Cooperative Ad Hoc Networks", IEEE SECON'2010, 2010, pp. 1-9.

[6] S. Kim, "Adaptive online power control scheme based on the evolutionary game theory", IET Communications,vol. 5, no. 18, 2011, pp. 2648 – 2655.

[7] Y. Eisenberg, and C. Logan, "Cooperative communications for improved throughput, range and covertness", Military Communications Conference, 2008, pp. 1-7.

[8] S. Mukhopadhyay, M. Jose, and D. Ghosh, "An Efficient Multiunit VCG Mechanism for the Ticket Booking Scheme of the J-League Football Tournament", IEEE International Conference on Industrial Informatics (INDIN), 2010, pp. 704-707.

Proposal for a NETCONF Interface for Virtual Networks in Open vSwitch Environments

Roberio Gomes Patricio*, Bruno Lopes Alcantara Batista*, Joaquim Celestino Junior* and Ahmed Patel†
* State University of Ceará
Fortaleza, Brazil
Emails: {roberio, bruno, celestino}@larces.uece.br
† University Kebangsaan Malaysia
Selangor Darul Ehsan, Malaysia
Email: whinchat2010@gmail.com

***Abstract*—The Internet is predominantly viewed as widely successful for existing users and service providers. But it suffers from ossification in the underlying infrastructure to exploit and scale to network virtualization for content providers and third-party hosting of cloud services through overlay networking by creating virtual ecosystems that enable and leverage new business opportunities. This paper proposes a Network Configuration Protocol (NETCONF) interface, modeled in YANG language, a data modeling language for networks. It provides standardized, simple and easy to use interfaces that facilitate the process of automating the creation of virtual networks using virtual switches, tested with Open vSwitch (OVS).**

Keywords*—*NETCONF, YANG, VLAN, distributed system.

I. Introduction

The increase of Internet data traffic and demands for faster and efficient network to accommodate social networking, big data and a stream of new virtualized cloud computing applications together with many different types of business opportunities is accelerating the tide of the next generation of the ubiquitous Internet. It scours the necessity of new technologies and new protocols that can support the creation of new kinds of networks to facilitate not only Internet evolution but also end user applications supported by core overlay network infrastructures. This has created a void that has become known as the ossification of the Internet.

New architectures and topologies have been proposed to resolve the Internet ossification problem [1], among them, Virtual LAN (VLAN) [2], Virtual Private Network (VPN) [3], and Overlays network [4] together with a series of underlying management protocols to support them. All of these enable the building of virtual network, that use the underlying native network infrastructure, which already exists to transform it into a new type of overlay network with new topologies and new management protocols. This allows the building of new computing based habitats, a kind of micro ecosystems, that are completely isolated from the nitty gritty of the underlying network infrastructures and their idiosyncrasies. It enables and leverages new business opportunities and strongly supports Cloud-Based Virtual Networks (CVN) [3]. CVNs are on their way that will transform the way ICT works, the way we humans and machines works.

To realize the benefits of virtualization, we need an architecture for network virtualization that encompasses the key players and providers such as players insides providers (PIPs) and service providers (SPs), virtual network providers (VNPs) for assembling virtual resources from one or multiple PIPs into a virtual network, and virtual network operators (VNOs) and virtual network providers (VNPs) for assembling virtual resources from one or multiple PIPs into desired virtual network. On the technical side, we need standardized interfaces between the players to automate the setup of virtual networks, ie, a common control plane. Moreover, we need ways in which each player can check if it is being provided with the service it is paying for (eg in terms of quality of service (QoS) and quality of experience (QoE).

These related initiatives contribute towards creating and using of the principles of Software Defined Networking (SDN), which represents a consolidation of the previously cited virtualization networking models that today are the object of study both in academia and within the major telecommunication, networking and ICT companies. They are promising to tide the next big application level networking wave.

When discussing about SDN and the initiatives of industrial applicability and related research, it is difficult to find anyone who does not cite Open vSwitch (OVS), a totally software based virtual switch that is able to provide a wide range of network services, such as among them VLAN and VPN. With OVS it is possible to create new virtual networks using a set of commands and a Command Line Interface (CLI) [5].

Within the last five years, end system virtualization, eg via Xen or VMware, has revamped server business. Router vendors such as Cisco and Juniper offer router virtualization, and existing techniques such as MPLS (Multiprotocol Label Switching) [6], GMPLS (Generalized MPLS) [7] and VPNs (Virtual Private Networks) [8] offer some coarse grained link virtualization. Overlays such as peer-to-peer (P2P) networks over the Internet (e.g., BitTorrent) can also be seen as a virtual network, but they suffer from a lack of sufficient isolation. VPNs (e.g., realized via MPLS), can also be seen as virtual networks. Open vSwitch is a production quality, multilayer virtual switch that attempts automation but is stuck with traditional forms of network interfaces. However, they suffer from a lack of node programmability using the latest software engineering tools.

Virtual networks can be tailored to meet a specific set of service provider and customer requirements that satisfy specific user groups under either public or private configuration management. While OVS as a tool goes some way to manage the configuration of theses environments, it lacks mechanisms of more versatile automation because, in particular, of the cumbersome and restrictive CLI that is out of phase with today's advanced software development and deployment technologies.

This paper proposes a Network Configuration Protocol (NETCONF) [9] interface, modeled in YANG language [10], a data modeling language for networks. It provides standardized, simple and easy to use interfaces that facilitate the process of automating the creation of virtual networks using virtual switches, tested with Open vSwitch (OVS).

The remainder of this paper is organized as follows: section II presents an overview of network configuration management using NETCONF and YANG; section III describes the related works done by other authors in this area; section IV describes the requirements and the modeling process used in this paper; section V presents a set of tests and results to evaluate and validate the proposed data model and section VI concludes the work with some conclusions and providing new ideas about future works.

II. Theorical Reference

A. NETCONF

The configuration management of a wide number of network elements and devices is still a major problem nowadays because of their complexities and vendor specific proprietary style of interactions. The mechanisms to retrieve and modify the configuration data are largely something specific of each device provider, and the configuration interfaces are difficulty to maintain and quite costly to achieve a high level of efficiency and reliability through automation especially when dealing with issues of maintenance and version control [11].

According to Case el al., [12] the NETCONF exceeds the deficiency of Simple Network Management Protocol (SNMP) and emerges as a promising approach to standardizing the mechanism of network management based in eXtensible Markup Language (XML). NETCONF provides a better configuration interface for network devices due to the effective use of technologies like XML and others. The philosophy behind NETCONF is the necessity of an interoperable programmable interface between the different network equipment vendors to manipulate the devices' configuration state of the entire network into a systematic whole [11].

B. YANG

The NETCONF protocol describes a communication model between network devices that need to be configured and managed. However, the specification of this protocol does not describe how the manipulated information in the data layer must be represented. This issue is taken up and addressed by the YANG data modeling language, that emerged from the working group called Netmod Standard Working Group (Netmod WG) [10] [11].

The fact of using XML messages, many other options of data representation emerged to work with NETCONF protocol, like XML and RelaxNG [2]. Despite the great power of expression of these languages and their wide adoption by the community, the Netmod WG chose to define their own data modelling language, aiming to have total control of it to achieve total independence from proprietary vendors. This is to avoid having to cater for specific formats and meanings that require data mapping transformation to achieve interoperability.

Thus, the YANG language as a data modeling language permits describing network elements. It covers not only information about the data configuration parameters and options, but allows handling data that describes the current state of the device and providing important and relevant data pertaining to network management. This goes way beyond just configuration management. It also allows tunneling the data and information to other aspects of network management such as accounting, security, performance and fault, in ISO parlance for network management, FCAPS [13].

Conceptually, according to McCloghrie et al. [14], YANG can be compared with Structure of Managed Information (SMI) [14], the language used by SNMP protocol [12] to define and construct network.

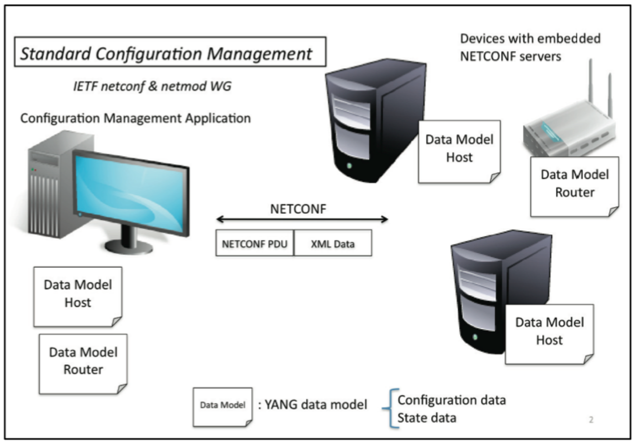


Fig. 1: The YANG and NETCONF integration

Management Information Base (MIB) that can be easily manipulated by the YANG data modeling language where such data are distributed and accessible only through NETCONF protocol. Figure 1 shows how to use the YANG language, its applicability and iteration with the NETCONF protocol.

III. Related Works

The Open Networking Foundation [15] discusses about the OpenvSwitch implementation and compares its performance with the Linux Bridge, which is the de facto reference implementation in the open source world with this purpose.

The drawback is that they do not offer sufficient in depth scenarios and examples of how to accommodate NETCONF or other essential forms of network management. One positive aspect of OpenvSwitch is, it offers centralized management control in a distributed environment, creation of VPNs and virtualized mobility between IP subnets.

Pfaff and Davie [5] proposed an OpenvSwitch's database management protocol based in JSON-RPC calls. They presented and discussed about insufficient details in the database schema and the Application Programming Interface (API) of calls to configuration management of instances of OVS, with their names, parameters and return types.

However, with the adoption of API based in JSON-RPC for integrating the activities of distributed systems and managing the configuration tasks of multiples instances in the absence of more robust and secure mechanisms for allowing transactional configurations, it actually ends up limiting the scope and horizons of virtualization applicability.

Furthermore, the granularity of exported services via JSON-RPC exposes the database in a way that violates the encapsulation of data and increases the demand of coupling of any clients who want to consume such services. It directly affects the evolution of the model by preventing inclusions of new requirements in the information model or updating the MIB.

In addition, none of the current APIs are able to present OVS with sufficient granularity in defining and accessing of their services from the point of view of the basic setup operations necessary for their proper functioning.

IV. THE YANG MODELING PROCESS

By using the OVS in real scenarios, we realized that we required an agile way of interaction with it. This was necessary to allow automated programmability of the configuration of OVS and its utilization in order to create virtual networks using VLANs as comprehensively as possible.

Initiating from a well known managed network environment concept, using the NETCONF protocol and data modeling language YANG, resulted in an innovative protocol that would offer a new communications interface in addition to the existing CLI and JSON-RPC. This also facilitated the OpenFlow specification [15], which states that the NETCONF protocol must be used for the most basic function of management and configuration in OpenFlow switches.

The following subsection shows the most important aspects of the proposed information model together with its major requirements. It also provides additional scenarios as visions of this model for a better understanding of its programmability and potential operational behavior. Beyond this, the data structures necessary to store the configuration data, the operations used via CLI for creating VLAN are also present in this model.

A. Multiple switch instances

We envisage to have inside of the one OVS process several other instances of virtual switches. This opens the possibilities to provisioning a specific instance for each client, starting from the same instance. In other words, a given OVS process shall actuate as several virtualized switches simultaneously, where each switch can be responsible for a distinct VLAN operation. To make this possible, for this to be functional ,we propose the following additions to our policy:

1) Strategy: The information model must anticipate/contemplate an object of a bridge type, which is contained within a collection called bridges as shown in Figure 2.

2) Vision: Figure 2 also shows the structure of the information model supporting multiple instances.

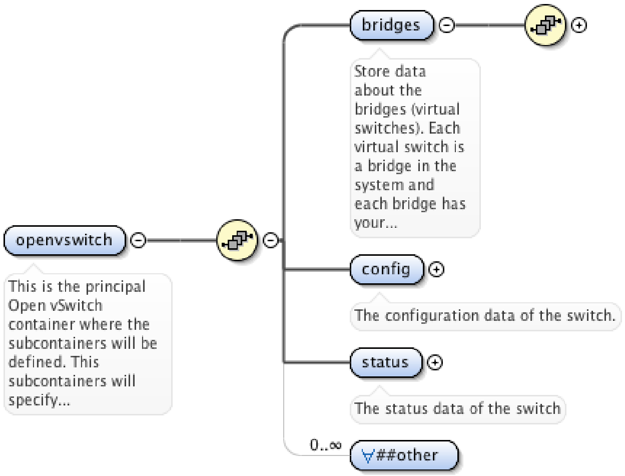

bridges
Store data about the bridges (virtual switches). Each virtual switch is a bridge in the system and each bridge has your...
openvswitch
This is the principal Open vSwitch container where the subcontainers will be defined. This subcontainers will specify...
config
The configuration data of the switch.
status
The status data of the switch
0..∞
∀##other

Fig. 2: The YANG and NETCONF integration

3) Normative Considerations: Following a document-based approach, the objects of the bridge type are grouped into a superior entity called bridges. This approach prevents such objects from being scattered inside the model. These may unnecessarily and considerably increase the NETCONF calls when the manager's function wants to obtain an overview of the whole, part, or only to access the instances of the OVS at once.

In this multiple switch instances case, using any other approach other than YANG would require unnecessary data normalization. Also using YANG over more traditional modeling approaches of MIBs in ASN1, which generally works with data structures that tie closely to standard Data Base Management Systems (DBMS), gives greater flexibility and independence of data types and DBMSs.

Tables would be costly in terms of operational efficiency and programmability. The YANG model is used as a container element to represent the bridges' objects and a list of elements for the bridges' objects. This approach is far more efficient than the traditional table approaches.

B. Multiply Ports in the same Switch Instance

Any virtual switch to add and configure more ports will be possible, with its respective network interfaces and VLAN tags when applicable. In this proposed policy the following are essential.

1) Strategy: The switch ports are to be mapped onto a port object, which are contained in a collection called ports.

2) Vision: Figure 3 shows the hierarchical root structure of the information model of how multiples ports can be supported in the same switch instance.

3) Normative Considerations: Group the port objects into a superior entity called ports. So it is possible to obtain information of all ports of a switch instance with a minimum of NETCONF calls.

Furthermore, the process of batch configuration of ports on a switch occurs in a much more rapid fashion, once all

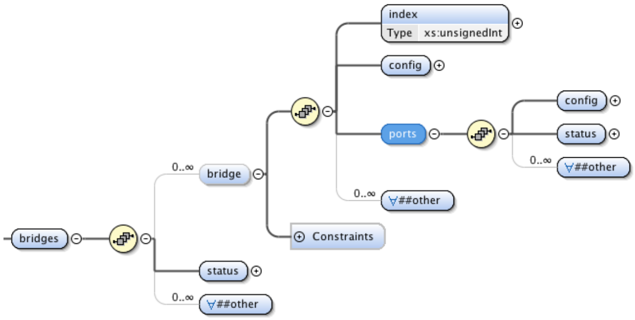


Fig. 3: The vision of multiples ports model

ports are under one entity. In terms of YANG, the container element represents the set of ports object and the list element represents the individual port object.

C. Configuration Data Must Not Be Exposed to Any User

The configuration data of a given entity, whether it is an OVS instance or a VLAN, is to be inaccessible in the model. The NETCONF protocol proposes a clear separation between all the configuration data and the state data.

Furthermore, it will be possible to access and retrieve statistical data of an object as well as configuration. In this proposed policy:

1) Strategy: The configuration data and the state data of a given element must be separated in distinct branches, opening the possibility to restrict the data access through defined rules on the NETCONF server.

2) Vision: Figure 4 shows a data model element containing two objects: config and status. The element called config contains the related data about the configuration of the device, while the element status stores the state data of the parent object, which can be used for statistical purpose.

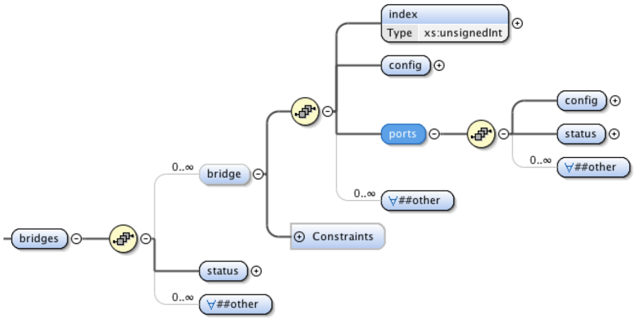


Fig. 4: The vision of config and status container model

3) Normative Considerations: The data separation is done through of two YANG containers. The definition of these containers is based on the kind of information that should be labeled with the YANG instructions, *config true* or *config false*.

Elements that have the YANG instruction *config true* must be stored in a config container and this data have read and write permissions. On the other hand, elements that have the YANG instruction *config false* must be stored in state container and this data have read-only permission. It is easy to associate access profiles to these elements in the Network Configuration Protocol (NETCONF) server and thus obtain more control over the access of these elements.

D. VLAN support

To create VLANs and associate them to the ports of a given switch instance, they must be properly identified by a tag or an ID, which should be provided during the creation of a virtual network. The policy for this case is:

1) Strategy: Each port of a given switch must be associated to an object called interface and can still be linked to a VLAN. This construction puts under one port all needed data for total VLAN support.

2) Vision: Figure 5 below shows the configuration of (1) above where the VLAN elements are connected in a given port of a given switch instance.

With this type of structure it is easy to configure a VLAN to a given port, since the data needed for this are easily accessible from the same branch in the tree of objects.

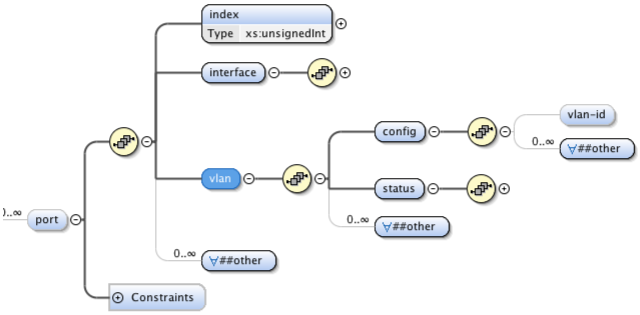


Fig. 5: The vision of VLAN support model

3) Normative Considerations: Here, two YANG containers which are nested and associated to a port container are used. The VLAN element is marked as optional, and may well be having the switch ports that are not being used in the construction of VLANs.

E. Well Defined API

A well defined API with a set of RPC operations could be used to manipulate the information model without the NETCONF client having the need to interact directly with the information model.

The granularity of these operations should not be much different to other user interfaces, such as, CLI to which the user is already accustomed. It also allows for making it easy to learn and take advantage of the previously obtained experiences working with other interfaces.

1) Strategy: Since the main operations are already supported by CLI, they become easily exportable to the NETCONF RPC call function. This can be executed in the context of an application of the network management system (NMS) in the standard programmable manner.

TABLE I: Current Supported Actions

Operation	Description	Params
add-bridge	This RPC call adds a bridge (virtual switch)	bridge name
del-bridge	This RPC call deletes a virtual swicth	bridge name
add-port	This RPC call adds a port in the specified switch	bridge name, interface name
del-port	This RPC call deletes a port in the specified switch	bridge name, interface name
add-vlan-port	This RPC call adds a port in the specified switch with a vlan tag	bridge name, interface name, vlan tag

Table I shows the currently supported operations by this model with their respective parameters.

V. Results

A YANG model, as described above, for the OVS is created to attend to the requirements already mentioned previously. We use the Yuma Server [16]. In our implementation of the NETCONF protocol, the Yuma Server is used to load the data model created by the YANG model by using the yangcli tool, that allows a NETCONF client to populate the data model in the associated MIB.

In the first test, we create two instances of OVS within the data model using the yangcli and as we can see in Listing 1. The model fulfills the requirements of multiple bridges. For reasons of simplicity, only the switches index are filled in the model, but in a real scenario all the other associated fields must be filled as well.

Listing 1: Multiple bridges on the NETCONF data model

```
rpc-reply{
 data{
  openvswitch{
   bridges{
    bridge 1 {
     index 1
    }
    bridge 2 {
     index 2
    }
   }
   status{
   }
  }
 }
}
```

The YANG model allows us to create several ports for a given switch instance, and according to the requirements, to have multiple ports in the same switch instance. Listing 2 shows a switch instance with two ports.

Listing 2: Multiple ports in the switch on the NETCONF data model

```
rpc-reply{
  data{
    openvswitch{
      bridges{
        bridge 1{
          index 1
          ports{
            port 1{
              index 1
              interface{
                config{
                  name eth0
                }
                status{
                  status up
                }
              }
            }
            port 2{
              index 2
              interface{
                config{
                  name eth1
                }
                status{
                  status up
                }
              }
            }
          }
        }
      }
    }
  }
}
```

In Listing 2, it can be observed that the configuration data and status data are separated into their respective containers, since the configuration data must be used exclusively by that process without exposure to any process.

In port 1 of switch 1 we have a container interface that has two sub containers, config and status, where the configuration data is within the config container (in this case the leaf name) and in the status container is the status data (in this case the leaf status).

Other requirement discussed in the previous section regarding VLAN support, follow the data modelling which allows determining the port of a given switch belonging to a VLAN. Listing 3 shows how a port is associated with a VLAN within the data model.

Listing 3: VLAN configuration in the switch port on the NETCONF data model

```
rpc-reply{
  data{
    openvswitch{
      bridges{
        bridge 1{
          index 1{
            ports{
              port 1{
                index 1
                interface{
                  config{
                    name eth0
                  }
                  status{
                    status up
                  }
                }
              }
              vlan{
                config{
                  valn-id 10
                }
                status{
                }
              }
            }
          }
        }
      }
    }
  }
}
```

The implemented model also has defined operations specified as requirements that make the appropriate API easily accessible in remote configuration using the NETCONF protocol. Besides implementing the data model in YANG, the API is written in C programming language that invokes the OVS operations.

Listing 4 shows an example of a virtual switch creation within the data model that is simple!.

Listing 4: Creating a switch and a port using API on the NETCONF data model

```
yangcli root@localhost> add-bridge bridge-name=sw01

RPC OK Reply 25 for session2:

yangcli root@localhost>
```

VI. Conclusion and Future works

This work proposed a NETCONF interface for the configuration management of virtual switches that should be used to create virtual networks based on VLAN.

The proposed information model and the currently supported operations were modeled in the YANG data modelling language, utilizing OVS as a reference of virtual switches. The interface currently supported by OVS may not be the most appropriate when one has to automate the process of configuring this type of switch, but it is sufficient to allow further work to be performed to overcome this deficiency.

In this paper, we listed the basic requirements that a virtual switch must meet to support virtual networks for VLAN. Each one of those requirements was attended to in the information model based on YANG and the solution was presented in a simple visual way using a XML Schema Definition (XSD) models. In addition, other policy considerations devolving into strategies used in the modeling were also presented.

The proposed information model was not intended to exhaust all the possibilities to take all the necessary requirements to a virtual switch, but, however to illustrate the information requirements and data modeling together with the set of the operations to realize in building virtualized networks based in VLAN with virtual switches. It is a principle that has huge potentials in creating the next generation of virtual networks supporting a variety of virtual content based applications in virtualized cloud computing. Simply put, it is putting up the ante for the next generation of the Internet that offers huge technical and business possibilities.

Further improvements can be implemented in this model, certainly the support of other approaches used to build virtual networks (such as OpenFlow for instance) are already being studied and can be easily incorporated in the proposed modeling.

Thus, using a NMS it will be possible to complete the configuration management of programmable virtual switches through a robust interface standards based that is consolidated every day as a great alternative to traditional managed network interfaces.

References

[1] G. N. Rouskas, “Tutorial on network virtiualization,” Presented at OFC/NFOEC, pp. 1393–1398, March 2012.

[2] N. M. K. Chowdhury and R. Boutaba, “A survey of network virtualization,” Comput. Netw., vol. 54, no. 5, pp. 862–876, Apr. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2009.10.017

[3] T. Choi, K. Nodir, T.-H. Lee, D. Kim, and J. Lee, “Autonomic management framework for cloud-based virtual networks.” in APNOMS. IEEE, 2011, pp. 1–7. [Online]. Available: http://dblp.uni-trier.de/db/conf/apnoms/apnoms2011.html#ChoiNLKL11

[4] O. vSwitch, “Open vSwitch: a open virtual switch,” [accessed April 2014], 2013. [Online]. Available: http://openvswitch.org/

[5] B. Pfaff, J. Pettit, K. Amidon, M. Casado, T. Koponen, and S. Shenker, “Extending networking into the virtualization layer,” in HotNets’09, 2009, pp. –1–1.

[6] E. Rosen, A. Viswanathan, and R. Callon, “Multiprotocol Label Switching Architecture,” RFC 3031 (Proposed Standard), Internet Engineering Task Force, Jan. 2001, updated by RFCs 6178, 6790. [Online]. Available: http://www.ietf.org/rfc/rfc3031.txt

[7] E. Mannie, “Generalized Multi-Protocol Label Switching (GMPLS) Architecture,” RFC 3945 (Proposed Standard), Internet Engineering Task Force, Oct. 2004, updated by RFC 6002. [Online]. Available: http://www.ietf.org/rfc/rfc3945.txt

[8] L. Andersson and T. Madsen, “Provider Provisioned Virtual Private Network (VPN) Terminology,” RFC 4026 (Informational), Internet Engineering Task Force, Mar. 2005. [Online]. Available: http://www.ietf.org/rfc/rfc4026.txt

[9] R. Enns, “NETCONF Configuration Protocol,” RFC 4741 (Proposed Standard), Internet Engineering Task Force, December 2006. [Online]. Available: http://www.ietf.org/rfc/rfc4741.txt

[10] Ietf, “RFC 6020: YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF),” Oct. 2010. [Online]. Available: http://www.ietf.org/rfc/rfc6020.txt

[11] H. Xu and D. Xiao, “Data modeling for netconf-based network management: Xml schema or yang,” pp. 561–564, 2008.

[12] J. D. Case, M. Fedor, M. L. Schoffstall, and J. Davin, “Simple network management protocol (snmp),” United States, 1990.

[13] H. Zimmermann, “Osi reference model–the iso model of architecture for open systems interconnection,” pp. 425–432, 1980.

[14] K. McCloghrie, D. Perkins, and J. Schoenwaelder, “Structure of Management Information Version 2 (SMIv2),” RFC 2578 (Standard), Internet Engineering Task Force, April 1999. [Online]. Available: http://www.ietf.org/rfc/rfc2578.txt

[15] O. N. Foundation, “OpenFlow ,” [accessed April 2014], 2013. [Online]. Available: https://www.opennetworking.org/index.php?option=com_content&view=category&layout=blog&id=57&Itemid=175&lang=en

[16] YumaWorks, “NETCONF ,” [accessed April 2014], Jan. 2014. [Online]. Available: https://www.yumaworks.com/netconfd-pro/netconf/

A new Unsupervised User Profiling Approach for Detecting Toll Fraud in VoIP Networks

Anton Wiens, Torsten Wiens and Michael Massoth
Department of Computer Science
Hochschule Darmstadt - University of Applied Science
Darmstadt, Germany
{anton.wiens | torsten.wiens | michael.massoth}@h-da.de

***Abstract*—Significant amounts of money are lost worldwide due to toll fraud attacks on telecom service providers or their customers. These attacks can be detected or prevented by a fraud detection system. Acquiring labeled data for the analysis of fraud cases is a major problem. This paper proposes an autonomous unsupervised user profiling approach for fraud detection using Call Detail Records (CDR) as data for the analysis and considers problems like random fluctuations in data. Two profiles for each user are used to measure user behavior in different time spans. The two profiles of every user are compared to each other, and changes in user behavior are measured. Describing the change in a numeric value allows checking for extreme changes and detecting fraud. For the detection of random events, a global profile is used. Two profiles are cumulating behavior information for all users, measuring global events in a reliable way. The approach provides low false positive rates. Also, recent fraud cases concerning Fritz!Box Voice over Internet Protocol (VoIP) hardware are analyzed and a detection approach based on this work is proposed.**

Keywords-Call Detail Record; Fraud Detection; autonomous unsupervised user profiling; VoIP.

I. INTRODUCTION

The Internet brought new possibilities for telecommunication (e.g., VoIP), and new communication channels have been created. But fraudsters also found their ways with those new possibilities. Fraudsters invade telephone systems and manipulate them to conduct expensive phone calls at the expense of the owner of the telephone system. The generated cost has to be paid by the users or the service provider most of the time, leading to large amounts of losses and even threatening the existence of small telecom service providers. Telecommunication fraud caused an annual cost in the hundreds of millions EUR at telecom service providers in the last years.

Communications Fraud Control Association (CFCA) reports losses of about 46 billion USD in 2013, an increase by 15% compared to 2011 [1]. But not only cost is a problem caused by fraud. Small providers may also suffer from reputation losses, causing customers to change the provider because of decreased trust and fear of repeated fraud attempts in the future.

The top three methods for telecommunication fraud were Subscription Fraud (subscribing for paid services), Private Branch Exchange (PBX)-Hacking and Identity Theft [1]. The top three types of fraud were Roaming (using stolen access in foreign countries), Wholesale (reselling of stolen user credentials) and Premium Rate Service fraud [1].

The German company “Deutsche Telekom” reported a huge success in the prevention of fraud cases with potential damages of about 200 million Euro, using an automated fraud detection system [2].

Recently, fraud cases were caused by security exploits in AVM Fritz!Box hardware, which is often used in Germany [3]. These fraud cases are analyzed in Section VII, and a detection approach based on the analysis is proposed.

The research project ‘’Trusted Telephony” at Hochschule Darmstadt pursues the goal to increase security and safety in VoIP telephony in cooperation with the German telecom service provider toplink GmbH. A key objective of the project is the development of a fraud detection system, consisting mainly of a software framework.

A huge problem for researching and developing a fraud detection system is the lack of labeled data. In labeled data, each record in the dataset is marked with the appropriate class for the dataset. In toll fraud detection, appropriate classes would be fraud and non-fraud. Labeling requires expertise and is a time consuming process. Because of this, labeled data is often not available, which is why autonomous and unsupervised techniques for fraud detection require less knowledge and personnel to maintain.

For this purpose, a technique has been developed that to work unsupervised and mostly autonomous. Full automation would require a final task for the software, the actual blocking of the customer or destination number. Due to the risk of automatically blocking a non-fraudulent customer or destination number, the approach proposed is autonomous except for this final task, which is done by the system administrator. Unsupervised means in this context that no explicitly generated training data is needed for this technique. It is based on an analysis of Call Detail Records (CDRs) and research on related work and applies user profiling, as well as assorted ideas from related work.

A CDR contains information about telephone calls, e.g., caller and callee, duration, and more. Because labeled data is often scarce, the developed method is designed to work without training a model with labeled data and to autonomously detect fraud in live operation, reducing the need and cost of administration by a staff member of the telecom service provider. The proposed method uses statistical profiles for each user for different time periods and continuously compares them in order to detect anomalies in the users’ behavior. Anomalies are distinguished as extreme changes in user behavior and are used to detect fraud. A Current Behavior Profile (CBP) describes the user behavior in the present, and a Past Behavior Profile (PBP) describes the behavior in the past. The profiles use statistical parameters (features) to describe the behavior in the time span of the profile. With a continuous comparison of those features of both profiles, an estimation of fraud or not fraud is made. This estimation is made by comparing the past profile with the present profile, analyzing extreme changes in behavior.

A. *Structure of the paper*

In Section II, related work is discussed. A definition of Call Detail Records is described in Section III. In Section IV, the reason for the usage of differential analysis and user profiling is explained as a basis for the concept following in Section V. Section VI describes an experimental evaluation of the proposed method with a first prototype implementation and its results. Finally, Section VII presents a conclusion on the proposed method and gives an outlook on future work.

II. Related Work

In related work, techniques for telecommunication fraud detection that do not require labeled data and are capable of autonomous detection (requiring no administration) are scarce. Much of the related work discusses methods that build profiles from labeled data, train machine learning algorithms and use the result for the evaluation of the data. As mentioned before, expert knowledge and a huge time effort is needed for this task.

Chandola, Banerjee and Kumar present a paper which is rich on information about anomaly detection in general and fraud detection, respectively intrusion detection for telecom networks [4]. As shown therein, most work is based on statistical approaches, neural networks and rule-based strategies.

In [5], two approaches are shown. One utilizes a neural network, trained with profiles and classifying profiles, and the other a statistical approach that has potential for automation. This is detailed more in [6] by the same authors. This method uses two profiles for each user. One is called “current user profile”, the other “user profile history”. The former describes the user behavior in the present, the latter in the past. For the description of the user behavior, so-called prototypes are used to group similar calls by time and duration of the call. Here, a prototype can be seen as a cluster, covering a certain range of values for time and duration of a call. Then, probabilities are calculated for these prototypes using the distribution of calls over the prototypes. A profile consists of probabilities for each prototype. The change in user behavior is measured using the Hellinger distance, which calculates differences between the probability distributions of both profiles.

This technique can potentially run autonomously, but still needs training for the prototypes. Also, the prototypes only use two attributes per call. Adding attributes exponentially increases the number of prototypes. The effects on performance and accuracy for an increased number of attributes are not specified in the paper. The idea to apply two profiles in different time spans to measure changes in user behavior has been a starting point for the method presented in the paper. In this work, the user profiles are built differently, the comparison of the profiles differs as well.

In [7], different user profiles have been evaluated in a combined neural network- and clustering-based technique to detect fraud. One profile type performed better than other profile types and therefore is used in Section V for the profiles describing the behavior of a user in different time spans. The profile consists of the following features: Standard deviation, maximum and mean values for the number of calls, the duration of calls and additionally the maximum cost per call.

In [8], an approach is proposed which combines identity authentication, key process monitoring and anomaly service traffic identification to detect and prevent fraud. There is scarce information on the implementation and no information about the results of the system, e.g., false and true positive rates.

In [9], a more potentially autonomous system for unlabeled data is proposed. It uses a rule-based approach to learn different types of so-called “monitors” that analyze user behavior and alarm the system’s administrator if fraudulent activity is detected. The system still needs templates and learned rules to create monitors, of which the templates need to be prepared and expert knowledge is needed.

Grosser et al. present in [10] an extension of the work in [5] by replacing the prototypes with a self-organizing map. The resulting system still lacks the ability to be autonomous.

In [11], a Bayesian Network is constructed for the detection of fraud in data. It uses the attributes Destination Country, Duration, Call Day and Call Type of a CDR.

As shown in [5], [12] uses a neural network trained with user profiles but different features to classify new profiles with. It results in a true positive rate of 90%, the false positive rate of 10% is quite high.

In [13], different attack patterns and possibilities for their detection are discussed.

In [14], many different approaches are shown: A neural network approach, a Bayesian network and an approach utilizing probability density estimations. All approaches apply user profiles with “…average and the standard deviation of the duration and the number of calls made during the day, maximum duration and number of calls per day during the observed time period…”.

Generally, a lot of work went into the analysis of machine learning techniques requiring training with labeled data which is hard to acquire. Only a few approaches allow to use unlabeled data. Most of them still require some sort of training, making automation hardly possible.

III. Call detail records

Each call of customers of toplink is routed through a dedicated voice routing system. Information about the call is recorded as a Call Detail Record (CDR) in text format in a file on a local hard disk drive. The data is then parsed with a parser developed in this project, and the necessary information is loaded into the project’s fraud detection framework. A CDR contains information about the connection and the call, e.g., IP addresses, trunk ID, start time, call duration, calling number, called number, customer ID, and much more. This data is analyzed for anomalies and potential fraud cases.

IV. ABSOLUTE OR DIFFERENTIAL ANALYSIS

In this section, the concepts of absolute and differential analysis are introduced. An absolute analysis examines a whole set of data, trying to identify fraud cases, but does not consider different types of user behavior. A call that may be treated as a fraud case for one user could be no fraud case for another user. For example, one user only makes long calls to his family at weekends and the other user only makes long calls to his family at workdays. If an absolute analysis considers long calls at workdays as fraud cases, the latter user will be considered as fraudulent, just because his normal behavior does not comply with the definition of normal behavior given by the other user. This problem can be avoided by looking at each user and his behavior differently, thus called differential analysis.

Differential analysis is preferred to absolute analysis in most of the related work, e.g., [5] [7] [10] [14]. The main argument is the ability of differential analysis to include the absolute analysis. In other words, a fraud case detected by an absolute analysis can also be found by a differential analysis, but a fraud case detected by a differential analysis cannot always be found by an absolute analysis [5].

User profiling is a differential analysis method, distinguishing the data by the users in the data. An analysis is then performed for each user on a smaller portion of the data, using only the data of the respective user.

For each user, profiles are constructed to measure the user behavior in a given time span from the user's data. A profile often consists of statistical features describing the user's behavior. For example, the mean duration of all calls or the mean number of calls in a given time span of the user data.

These user profiles are then used for training machine learning or other techniques to detect fraud cases by the values of each profile.

V. BASIC CONCEPT OF USER PROFILING APPROACH

Without labeled data, only few machine learning techniques may be used for fraud detection. Supervised techniques, e.g., a neural network as in [14], need a training phase with prepared, labeled data.

User profiling with statistical methods is therefore used as an unsupervised and autonomous approach. Two user profiles are generated for each user, describing user behavior in two different time spans, allowing for the detection of anomalous changes in user behavior by the comparison of the user's behavior in these two time spans. The user behavior in both profiles is described by the same features. The following sections are giving a more detailed description of the proposed method.

A. *Constructing user profiles*

For each user, two user profiles exist that represent the present and past behavior in specified time spans. The profile describing the past is called Past Behavior Profile (PBP), and the one describing the present is called Current Behavior Profile (CBP). Each profile uses features, calculated from CDR data, to describe the user behavior in its time span.

1) *Features*

Features describe different aspects of a user's behavior. In the profiles, the feature vector shown in Table 1 was used:

TABLE I. FEATURE VECTOR USED FOR USER PROFILES [7]

Max Calls	Max Duration	Max Costs	Mean Calls	Mean Duration	Std Calls	Std Duration

These are the maximum values (Max) for calls per hour (Calls), the duration of a call and the cost of a call, the mean value (Mean) and standard deviation (Std) for the same CDR information, except the cost.

For those features, the start-time, duration and cost information of a CDR are needed. The cost of a call is depending on the user agreement and is not given in a CDR. Therefore, an approximation of costs for a CDR was made, based on country code, number type (mobile or fixed-line) and duration.

These features were used because they delivered the best results in [7]. Many works use standard deviation and mean values of the number of calls and the duration of a call to describe the user's behavior. Some works also differentiate them into national, international or mobile [10] [7] [14].

2) *Profile Time Span*

Each profile P has a length l_P. The PBP additionally has an offset $d \neq 0$, describing the difference in time between the present and the PBP time span (see Figure 1). For a CDR to be included in a profile, it needs to meet the following rules (1) and (2) for the corresponding profile:

$$T_{cdr} < T_n - d \qquad (1)$$

$$T_{cdr} \geq T_n - (l_P + d) \qquad (2)$$

T_n is the present (n) time, and $T_{cdr}$is the time of the CDR. If a CDR meets these two rules, it is included in the features of the corresponding profile.

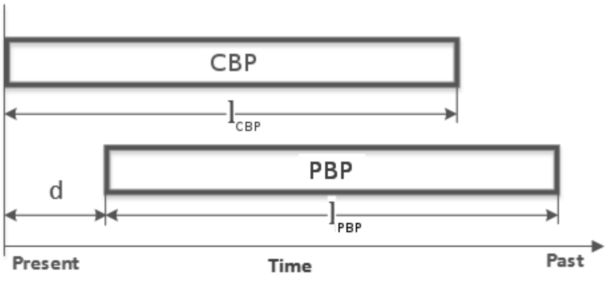


Figure 1. Profile time spans and offset (CBP = Current Behavior profile; PBP = Past Behavior Profile)

The length (time span) of the profiles and the offset are very important parameters for the detection. The longer a profile is, the more CDRs are represented inside a profile and the statistics have more accuracy and less fluctuations. At the same time, the effects of single fraudulent CDRs become statistically more irrelevant and thus harder to detect. The offset is important for finding fraudulent CDRs that can only be found in groups. It decides how long it takes for a yet undetected fraud CDR to be included in the PBP and therefore make it more unlikely to be found. The length of the offset also affects fluctuations when comparing both profiles. A higher offset causes higher fluctuations, a lower offset causes lower fluctuations likewise.

An optimal tradeoff between the length of the profiles and the offset between profiles needs to be found for best results.

3) *Filling Profiles*

At first, the profiles need to get filled up for the method to be able to calculate meaningful features. Once the profile contains CDRs for its entire time span, the features can be calculated and used for further analysis. This means that the method has a determined training time for accumulating CDRs that is autonomously done without administration by personnel. In the following, a profile that has been filled up once is called *ready*.

B. *Measuring change in user behavior*

Once the profiles of a user are *ready*, the change of behavior measured by the profiles can be calculated. This is done by calculating the relative ratio R_F between each feature F of both profiles (PBP and CBP) by (3):

$$\forall F : R_F = \begin{cases} \left(1 - \left(\frac{F_{PBP}}{F_{CBP}}\right)\right), & F_{PBP} \leq F_{CBP} \\ \left(1 - \left(\frac{F_{CBP}}{F_{PBP}}\right)\right), & else \end{cases} \quad (3)$$

This results in a ratio R_F for each feature F, describing the change in behavior for that feature. Each R_F has a range of -1 to 1, with -1 as a maximum decrease and 1 as a maximum increase in behavior measured by that feature.

A ratio R_F for a feature F gives a relative value to the past behavior. It is relative because the severity of a change in user behavior is always relative to the past behavior of the user.

1) *Empty profile*

In the case that a user did not make calls for a time span greater than the span of all user specific profiles, one of the profiles of a user can run empty. Once a profile is empty, the calculation of the features is not possible, because they attain a value of zero. Comparing a non-empty profile with an empty profile will result in infinite ratios for the features, allowing for detection of fraud where there is none (e.g., when the PBP is empty and the CBP is not empty). Instead of letting the profile run empty, the last CDR in a profile that is about to become empty is not removed. This prevents the features from getting zero values and keeps user specific information for fraud detection. Setting the features to a standard value would disregard user specific behavior and is therefore not done.

2) *Features accepting zero*

Features like standard deviation can attain a value of zero, even if the profile is not empty. For example, the standard deviation of the duration attains zero, if all calls in the profile have the same duration. Like in an empty profile, zero values are a problem for calculating the ratios. Therefore, a value ε (depending on the range of the specific feature) is added to the affected feature in both profiles.

C. *Detecting fraud*

For this approach, fraud cases are to be distinguished by extreme changes in user behavior described by each feature. Thus, for each ratio R_F of a feature F, a limit L_F is introduced. Each ratio R_F is therefore checked if its limit L_F is exceeded, and the number (n) of exceeded limits is checked against an additional limit L_E (E for exceedings). If the limit L_E is exceeded, the CDR is labeled as fraudulent and as non-fraudulent otherwise. The procedure can be described as follows:

1. Set $n := 0$
2. $\forall R_F \in R: (R_F > L_F) \rightarrow (n = n + 1)$
3. $result = \begin{cases} fraud, & n > \mathrm{L_E} \\ normal, & else \end{cases}$

Once a CDR in the CBP is labeled as fraudulent, it is to be excluded from inclusion into the PBP. This prevents the PBP from including fraud cases and obscuring potential follow-ups of fraudulent CDRs. This is the first approach chosen for a first experiment. Other approaches for detection using the ratios are discussed in future work.

D. *Unexpected fluctuations*

Many fluctuations in data and ratios, like weekends and holidays, can be predicted and adjusted for. But there are also fluctuations caused by random events inside the telecom service provider’s network, e.g., network, hardware or other failures.

Those fluctuations are hard to predict using user profiles. The idea is to use the relation between absolute and differential analysis. If it is a fluctuation caused by the specific user, the fluctuation is not seen in an absolute analysis. If the fluctuation is global, it will affect all users and will be seen for specific users, too. Therefore, the accumulated behavior of all users has to be measured to detect this kind of fluctuation.

Because the functionality to measure user behavior has already been defined, it can be reused to measure the accumulated user behavior. A global version of a CBP and a PBP is needed for all users. Ratios are calculated the same way as in user profiles. In this case, the ratios are not used for fraud detection, because the source of the fraud cannot be detected by creating profiles for all users. The ratios are used to be included in the user specific ratios for finding the global fluctuations and removing them from user fluctuations.

The inverse ratios of the global profiles are taken to the power of g and are multiplied with the corresponding ratio of a specific user profile as in (4):

$$newratio = (\mathbf{1} - globalratio)^g \cdot userratio \quad (4)$$

An appropriate value for g is determined in Section VI. Both ratios have the same scaling and global ratio that describes the change for the user ratio that is still normal.

Therefore, the inverse is multiplied by the user ratio. Because the global ratio is much more stable with more samples, it is taken to the power of g. g is dependent on the scaling of $globalratio$ and not on $userratio$.

E. *Low usage users*

An analysis of the data revealed that on average, each user only makes 6-7 outgoing calls per day. About 47% of the users only make 2 calls per day on average. That means a lot of users — and therefore user profiles — include low amounts of calls. Hence, only few samples are available for calculating the statistics, making the statistics inaccurate. A way to handle those fluctuations is to scale the calculated ratios for the user by the number of samples inside the profiles. For the creation of a scaling function$S(x)$, the dependencies of the number of calls in the profiles and the ratios needed to be analyzed. The analysis and the function are described in more detail in Section VI.

Before and after scaling a ratio, it needs to be converted to linear space with (5).

$$S(x, y) = 1 - \frac{1}{\left(\left(\frac{1}{1-y} - 1\right) * S(x)\right) + 1} \tag{5}$$

x is the number of calls in the PBP, and y is the ratio to be scaled. The part $\left(\frac{1}{1-y} - 1\right)$scales the ratio into linear space, and $1 - \frac{1}{(\ldots)+1}$ reverts it back to the previous space. A full overview of all components and their relationships is shown in Figure 2.

VI. Experimental Results

This section describes the test of a prototype implementation in an experiment. The implementation has been done in Java for an existing fraud detection framework of the research project. The data used for the experiment has been generated by a live environment, recorded by the VoIP switching device. The data consists of 76,326 cost impending calls and spans over a time of one month. It has been anonymized in accordance to the German Federal Law on Data Protection.

For the experiment, the whole data set was used, as the system trains on live data with the assumption that fraud cases are rare enough that the profiles can initially be trained by themselves without greater risks of being manipulated by fraud cases. Assuming the contrary is true and the first data set is containing fraudulent CDRs, the impact would only be that no fraud cases are detected until the fraudulent CDRs are no longer used for the PBP.

For the experiment, profiles of a week’s length and with an offset (d) of one day for the PBP are used. In a first run, all occurring ratios are recorded to calculate limits for the ratios, to analyze the parameters for the scaling function and to integrate the global ratios into user profiles. In a second run, the limits were applied and the fraud detection component was enabled.

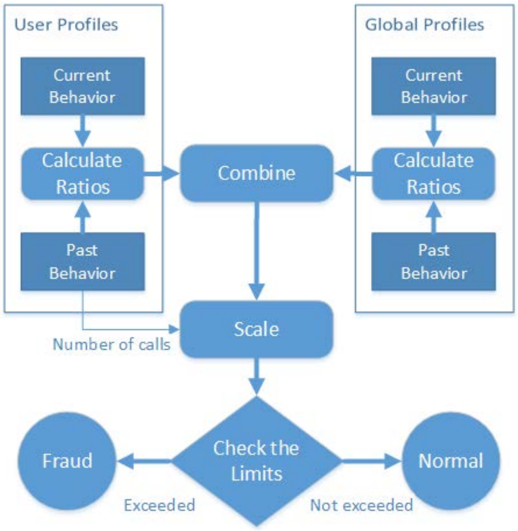


Figure 2. Overview of the components and their relationships

A. *First results*

For the first results, without incorporating the global profiles and the scaling function, the false positive rates (FPR) for different limits were measured. The false positive rate is a very important measure that indirectly determines the expenses due to inefficiency, because administrators need to look at false positives.

TABLE II. First results of FPR without global profiles and scaling for different limits

Limit for all ratios	Limit for exceedings	FPR
0.25	>0	0.2142
0.25	>1	0.1274
0.5	>0	0.0685
0.5	>1	0.0444
0.75	>0	0.0211
0.75	>1	0.0145

Table II shows empirically tested limits for ratios and the number of exceedings. The FPR has been measured from 50,893 samples, where the profiles were *ready*. The limits and the resulting FPRs will be used for comparison with results of the incorporations of global profiles and the scaling function for low usage.

B. *Global profiles*

For the global profiles, the same length and offset was used, because the ratios can be compared better if the parameters are similar. The number of calls was used as the only feature for the global profiles. For the parameter g for scaling the global ratio, see (4), a test value of 1 was used.

Figure 3 shows the ratios measured for the given data, chronologically sorted. It shows negative ratios during the Christmas holidays in Germany, successfully measuring its effects on the ratios and it can be used to remove those effects from single user behavior. Also, this figure shows when the profiles became *ready*.

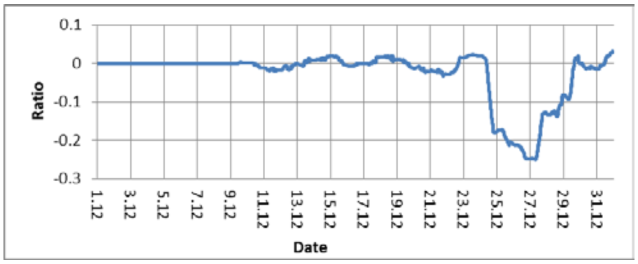


Figure 3. Ratios for number of calls for the whole data in global profiles

The incorporation into profiles of a week's length showed no significant improvements in the FPRs. On the other hand, a small scale test of profiles with a day's length showed very good results in removing weekend fluctuations from the profiles. Figure 4 depicts an example for day-length profiles.

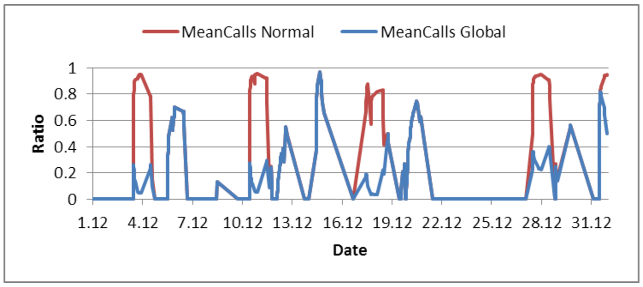


Figure 4. Example incorporation of global ratio into a day length user profile for feature MeanCalls

The figure shows two curves, MeanCalls Normal showing the ratios of the feature MeanCalls without correction by global profiles and MeanCalls Global with correction by global profiles.

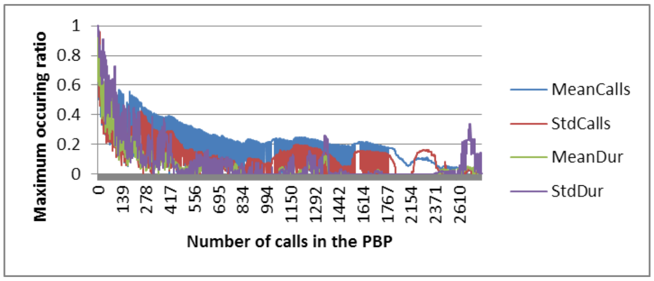


Figure 5. Example for the dependency of max values of the features MeanCalls, StdCalls, MeanDur and StdDur to the number of calls

C. Scaling for low usage

To find an appropriate scaling function, the dependency of the number of calls to the maximum occurring ratios was analyzed. Figure 5 shows an example for four features. It depicts how a low number of samples/calls in a profile can affect the ratios. Therefore, a scaling function was created that scaled the ratios from 0 to 70 calls.

For the scaling function, a simple parable of the form $y = (ax)^2 + b$ was chosen after testing different curves, because it corresponds well to the curve in Figure 5. Using the coefficients $a = \frac{1}{67.1}$ and $b = 0.2$, the scaling begins at 0.2 with 0 calls and ends at 1 with 60 calls with a slight increase. Because about 47% of users only conduct about two calls per day, the scaling function greatly improved the FPRs, as shown in Table III.

TABLE III. CHANGES IN FPR WITH INCORPORATION OF THE SCALING FUNCTION

Limit for all ratios	Limit for exceedings	Old FPR	New FPR	Change in %
0.25	>0	0.2139	0.1684	-21,27%
0.25	>1	0.1272	0.0939	-26,17%
0.5	>0	0.0683	0.0491	-28,11%
0.5	>1	0.0443	0.0290	-34,53%
0.75	>0	0.0211	0.0136	-35,54%
0.75	>1	0.0145	0.0083	-42,75%

D. Determination of limits

The best way to determine the limits is to optimize the ratio of true positive rate to false positive rate. However, this requires labeled data to be possible. Because of the lack of labeled data, the limits were determined by measuring the 99.5% quantile of all occurring ratios for each feature. The ratios are presented in Table IV. Using these limits, the measured FPR is 1.87%.

TABLE IV. LIMITS FOR FEATURES (99.5% QUANTILE)

Feature	Limit
MaxCalls	0.8247
MaxDur	0.6692
MeanCalls	0.7512
StdCalls	0.8270
MeanDur	0.2985
StdDur	0.5400
MaxCost	0.7387
Mean	0.3835

E. Results

Of the 50,893 analyzed cost impending calls, 1.87% were measured as false positives. Through empirical inspection of the false positives, two users were found with an exceptionally strange behavior pattern. The duration of calls and the number of calls per second was the same in about 200 calls, which is very suspicious. After consultation with toplink GmbH, those calls were considered fraud cases. This shows that the presented approach can detect false positives and reduce the FPR to 1.22%, but does not provide a true positive rate for a decent comparison with related work. Still 90.23% of the fraudulent calls found in these two users were marked as fraud by the proposed approach. Compared to the approach proposed in [6], which also proposes a statistical, unsupervised method, the approach of this paper has a lower FPR (1.22% to 4.0%). Compared to other supervised techniques, like [12] (with 50% TPR and 0.3% FPR) or [14] (two approaches with 70% and 80% TPR and 0% FPR for both), the proposed approach has a good TPR and FPR and needs no effort for preparing supervised training data.

VII. CONCLUSION AND FUTURE WORK

This approach allows the detection of fraud cases using unlabeled data and needs no maintenance by an administrator concerning data for training. Only administration for a final decision on positively identified fraud cases is needed. It is not complex and highly modifiable. It has a low false positive rate and allows detection of fraud cases with an estimated high true positive rate. The scaling for users with low usage rates still needs adjustment, and more profiles need to be tested with other features.

In the future, an autonomous limit adaptation is scheduled to be developed, making manual calculation of limits for the ratios obsolete, and making this approach even more autonomous and efficient. Because of the adapted limits, scaling the ratios for users with low activity is not needed anymore. Also, the limits will provide a more stable FPR for seasonal and user dependent behavior changes.

Furthermore, automation of unsupervised techniques requiring training could be possible by using a sliding window approach on the data consisting of present and past profile values used for training and testing. A support vector machine (SVM) is foreseen to be utilized, possibly including a feature preparation method, as proposed in [15]. This could also be seen as a test for using the results of the proposed approach as an input for supervised techniques. As mentioned in Section II, only few works are available that use techniques capable of being unsupervised and autonomous. Most approaches use techniques requiring training with labeled data and have no potential for automation.

Recent fraud cases, allowed by security exploits in Fritz!Box hardware, showed a repeating pattern in fraud attacks. These attacks utilized the hardware of many customers to call a single fee-based service or number, obscuring the attack by generating only few calls from each customer. A custom version of the approach proposed in this paper will be able to detect such attacks by profiling not the customers, but the destination of the calls. Such a profiling would record the amount of call attempts by different customers to a specific destination and detect extreme changes, enabling detection of fraud cases.

ACKNOWLEDGMENT

The state of Hessen, Germany is supporting this work with funds from the development program LOEWE. toplink GmbH from Darmstadt, Germany is cooperating with Hochschule Darmstadt on this project, providing essential information for fraud detection, e.g., the test data mentioned herein.

REFERENCES

[1] Communications Fraud Control Association, "Global Fraud Loss Survey," October 2013. [Online]. Available: http://www.cfca.org/pdf/survey/CFCA2013GlobalFraudLossSurvey-pressrelease.pdf. [Accessed 23 04 2014].

[2] heise online, "Bericht: Deutsche Telekom wertet Verbindungsdaten sämtlicher Telefonate aus," 10 08 2013. [Online]. Available: http://www.heise.de/newsticker/meldung/Bericht-Deutsche-Telekom-wertet-Verbindungsdaten-saemtlicher-Telefonate-aus-1933436.html. [Accessed 23 04 2014].

[3] AVM GmbH, 06 02 2014. [Online]. Available: https://www.avm.de/de/News/artikel/2014/sicherheitshinweis_telefonmissbrauch.html. [Accessed 23 04 2014].

[4] V. Chandola, A. Banerjee and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surv., vol. 41, no. 3, p. 15:1-15:58, 2009.

[5] P. Burge, J. Shawe-Taylor, C. Cooke, Y. Moreau, B. Preneel and C. Stoermann, "Fraud detection and management in mobile telecommunications networks," in European Conference on Security and Detection, 1997, pp. 91-96.

[6] P. Burge and J. Shawe-Taylor, "Detecting Cellular Fraud Using Adaptive Prototypes," in Proceedings AAAI-97 Workshop on AI Approaches to Fraud Detection and Risk Management, AAAI Press, 1997, pp. 9-13.

[7] C. S. Hilas and P. A. Mastorocostas, "An application of supervised and unsupervised learning approaches to telecommunications fraud detection," Knowledge-Based Systems, vol. 21, no. 7, pp. 721-726, 2008.

[8] Dai Fei Guo, Ai-Fen Sui and Lei Shi, "Billing attack detection and prevention in mobile communication network," in IEEE 13th International Conference on Communication Technology, 2011, pp. 687-691.

[9] T. Fawcett and F. Provost, "Adaptive Fraud Detection," Data Mining and Knowledge Discovery, vol. 1, no. 3, pp. 291-316, 1997.

[10] H. Grosser, P. Britos and R. García-Martínez, "Detecting fraud in mobile telephony using neural networks," in Proceedings of the 18th international conference on Innovations in Applied Artificial Intelligence, Bari, Italy, Springer-Verlag, 2005, pp. 613-615.

[11] T. Kapourniotis, T. Dagiuklas, G. Polyzos and P. Alefragkis, "Scam and fraud detection in VoIP Networks: Analysis and countermeasures using user profiling," in 50th FITCE Congress , 2011, pp. 1-5.

[12] Y. Moreau, H. Verrelst and J. Vandewalle, "Detection of Mobile Phone Fraud Using Supervised Neural Networks: A First Prototype," in Proceedings of the 7th International Conference on Artificial Neural Networks, Springer-Verlag, 1997, pp. 1065-1070.

[13] M. Nassar, S. Niccolini, R. State and T. Ewald, "Holistic VoIP intrusion detection and prevention system," in Proceedings of the 1st international conference on Principles, systems and applications of IP telecommunications, New York City, New York, ACM, 2007, pp. 1-9.

[14] M. Taniguchi, M. Haft, J. Hollmen and V. Tresp, "Fraud detection in communication networks using neural and probabilistic methods," in Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, 1998, pp. 1241-1244.

[15] D. Wang, Q.-y. Wang, S.-y. Zhan, F.-x. Li and D.-z. Wang, "A feature extraction method for fraud detection in mobile communication networks," in Fifth World Congress on Intelligent Control and Automation, vol. 2, 2004, pp. 1853-1856.

Towards a Carrier Grade SDN Controller: Integrating OpenFlow With Telecom Services

Caio Ferreira[†], Natal Neto[†], Alex Mota[†], Luiz C. Theodoro[†], Flávio de Oliveira Silva[*†],
João Henrique de Souza Pereira[*], Augusto Neto[‡§],Daniel Corujo[‡],
Carlos Guimarães[‡], Pedro Frosi Rosa[†], Sergio Takeo Kofuji[*] and Rui Aguiar[‡]
[*]Polytechnic School, University of São Paulo - Brazil
São Paulo, São Paulo, 05424-970
Email: flavio@pad.lsi.usp.br, joaohs@usp.br, kofuji@pad.lsi.usp.br
[†]Faculty of Computing, Federal University of Uberlândia - Brazil
Uberlândia, Minas Gerais, 38400-902
Email: {flavio, frosi}@facom.ufu.br, {caiocf,natal,alex,lclaudio}@algartelecom.com.br
[‡]Instituto de Telecomunicações, Universidade de Aveiro - Portugal
Aveiro, Portugal, 3810-193
Email: {dcorujo,carlos.guimaraes,ruilaa}@ua.pt
[§]Dept. de Informática e Matemática Aplicada(DIMAp), Universidade Federal do Rio Grande do Norte - Brazil
Natal, Rio Grande do Norte, 59078-970
Email: augusto@dimap.ufrn.br

***Abstract*—Software-Defined Networking (SDN) essentially decouples the hardware from the software that controls it. Currently, some SDN abstractions are materialized by OpenFlow and several OpenFlow controllers, based on different programming paradigms and architectures, are available. Usually, these controllers are bundled with some sample applications that enables the construction of new ones by using their particular way. However, these applications focus on specific services being tightly coupled with the switch behavior. In this scenario, SDN community is working to build a SDN control layer that meets carrier grade requirements such as throughput, availability and scalability. This work proposes a new SDN controller architecture that is integrated with a carrier grade service level execution environment, based on Service Logic Execution Environment (SLEE) architecture, defined under the Java APIs for Integrated Networks (JAIN) initiative. The proposed approach extends SDN based services by integrating OpenFlow with several network resources and communication protocols providing a cross layer platform that can satisfy these telecom operators requirements.**

Keywords–Software-Defined Networking; Carrier Grade; Controller; Telecommunications.

I. Introduction

Software-Defined Networking (SDN) [1], [2] is a promising networking technology since it has the potential to enable innovation and also to give the network operators more control of their infrastructure. SDN market is expected to reach thirty five billion dollars by 2018 [3], [4]. Essentially, SDN decouples the forwarding plane from the control plane. Currently, Openflow [5] materializes some concepts of SDN. In such approach, a signaling protocol is defined for the networking control plane, which enables controllers to orchestrate OpenFlow-compliant network devices (e.g., switch and wireless access points) in a programmable way by a controller.

The OpenFlow-enabled SDN is a key contribution that is propelling the networking research community towards the definition of the Future Internet, allowing the use and evaluation of innovating mechanisms for both network control and data transport. The list of Future Internet mechanisms include, among others, innovative approaches for routing, mobility, Quality of Service/Experience control, optical resource control. Future Internet Testbeds Experimentation Between Brazil and Europe (FIBRE) [6], Global Environment for Network Innovations (GENI) [7], OpenFlow in Europe: Linking Infrastructure and Applications (OFELIA) [8] and Abstraction Layer for Implementation of Extensions in Programmable Networks (ALIEN) [9] are examples of research projects across the world that are envisaging to spread the use of SDN by enabling the experimentation on top of an OpenFlow-enabled infrastructure of some of these innovate mechanisms.

The high demands for improving more and more the reliable aspects of their network systems and also the ability to take a complete control of their infrastructure, allowing customization and optimization and thus reducing overall capital and operational costs recently sparked the interests of telecom companies in exploring the use of OpenFlow-enabled SDN.

However, SDN poses some research challenges [10] to the scientific community. A key challenge is related with the controllers that should meet carrier grade requirements [11] that encompasses high availability, scalability, high performance, reliability, fault tolerance and manageability in order to foster the SDN adoption in mission critical environments, such as the ones handled by telecom operators.

Currently, several telecom operators have services [12] deployed on top of a mature platform, known as JAIN SLEE. The JAIN [13] is a set of APIs dedicated to creating voice and data convergent services. The JAIN SLEE [14] is a component model that supports the deployment of event driven applications that requires carrier grade requirements [15], [16], [17]. JAIN SLEE is available as commercial products, such as Rhino [18], and also as open source platform, such as Mobicents [19].

At this moment, the SDN community is still in pursuit of a carrier grade SDN control layer that is suitable for the demands of such environments. This work presents a contribution to this research by using an approach that differs from efforts currently undertaken by the SDN community in relation to this quest. This approach consists of constructing a controller layer that uses the component model defined by JAIN SLEE.

The remainder of this paper is structured as follows: Section II describes related projects and presents the ecosystem of OpenFlow controllers currently available. Section III presents the JAIN Slee component model and the components created as the basis for the carrier grade SDN controller that enables a cross layer approach by integrating OpenFlow with protocols used by telecom's voice and data services. Section IV presents an application scenario where the proposed approach was used to integrate OpenFlow and the Multimedia Independent Handover (MIH) protocols [20] and finally, in Section V, we present some concluding remarks and future work.

II. OpenFlow Controllers Ecosystem

Currently, the literature notices the availability of a number of OpenFlow controllers. In essence, they differentiate each other by programming paradigms or focusing on applications. For instance, Nox [21], [22] was the first designed OpenFlow controller deployed based on C/C++ programming language. Java based version of controllers were also created such as Beacon [23] and Maestro [24]. POX [22] is a Python based version of the controller, which offers a simplified programming interface enabling rapid deployment of new network applications. Ryu [25] is also based on Python and supports OpenFlow versions 1.0, 1.2 and 1.3. Trema [26] is a controller based on Ruby and FlowER is an Erlang based Openflow Controller [27].

These controllers came with sample code that shows how to create new network applications by using each controller proposed approach. A classical example of such applications is a *Learning Switch*. Usually, these applications offer low-level services and the network developer is responsible to build new ones according to particular requirements. While they are suitable as a starting point to use SDN concept, these controllers are not enough to achieve reliability and performance carrier grade networks demands [28], [29]. Such demands are not related to the telecommunication capabilities of the network entities but with aspects such as high availability, scalability, high performance, reliability and resiliency. The open issues of the aforementioned OpenFlow controller guided current efforts, including FloodLight [30], Onix [31] and more recently, OpenDaylight project [32] and ONOS [33].

FloodLight was created as a fork of Beacon, where the focus is to build a commercial controller with enterprise class. The open source version does not offer resiliency or scalability such as the commercial version, called Big Network Controller [34], which is based on a clustered servers in a HA deployment and uses in its core FloodLight.

Motivated by the inability in satisfying neither reliability nor scalability, the NOX creators proposed Onix, which is a distributed system over the network control plane that offers a global view of the network and defines an API that can use this information to build new control plane services and addresses such requirements. Onix was the basis for the software offered by Nicira [35] and its approach is used by the Modern SDN Stack [36] project.

The OpenDaylight project [32] aims to create a common architecture that can be exploited by the industry to create new and innovative services which use SDN abstractions. This architecture comprises several layers, one of them being the Service Abstraction Layer (SAL) that could interact with different protocols that would be exposed by plug-ins. Another layer is the Controller Platform [32] that controls the network devices, such as routers and switches, and defines a common API that will be used by the upper layer applications. One of the objectives of OpenDaylight project is to create a carrier grade architecture [37].

Another open source controller that currently is under development is ONOS (Open Network Operation System) [33], which aims at providing an architecture focused on fault tolerance and distribution of the state in various controllers, and providing a graphical high-level abstraction of the network status. According to ON.LAB, these features make ONOS a good alternative for service providers and also large WAN operators. A prototype was presented by OnLab at ONS 2013 and also at 2014 indicating that ONOS is still on the way to reach its goals [38].

Big Network Controller and also Onix are not available as open source and thus the SDN community is not able to run experiments using these particular solutions. OpenDayLight project and also ONOS have a road ahead.

Moreover, all these SDN controllers do not offer by default an integration with other signaling approaches used by telecommunications operators, such as: Session Initiation Protocol (SIP) [39]; DIAMETER [40]; Extensible Messaging and Presence Protocol (XMPP) [41]; Media Gateway Control Protocol (MGCP) [42]; among others.

This scenario indicates some open issues and this work contributes to bridge this gap.

III. Carrier Grade OpenFlow Controller

The network infrastructure itself has no value. The value is in the applications and services that can be created on top of this infrastructure. SDN abstractions enable the deployment of new and innovative services and even completely new network architectures [43]. However, these abstractions are being deployed at this moment and the SDN community is still in pursuit of a carrier grade platform.

The work presented here is deployed on top of the JAIN [13]. The JAIN is a set of APIs [44] dedicated to creating voice and data convergent services. The goal of these APIs is to abstract the underlying network, so those services can be developed independently of network technology. This approach couples with SDN abstractions.

The JAIN SLEE [14] defines a component model that supports event driven applications suitable for carrier-grade environment concerned with requirements such as high throughput, low latency, scalability [15] and availability [17]. Currently, several telecom operators have services deployed by using JAIN SLEE. JAIN SLEE is available as commercial products,

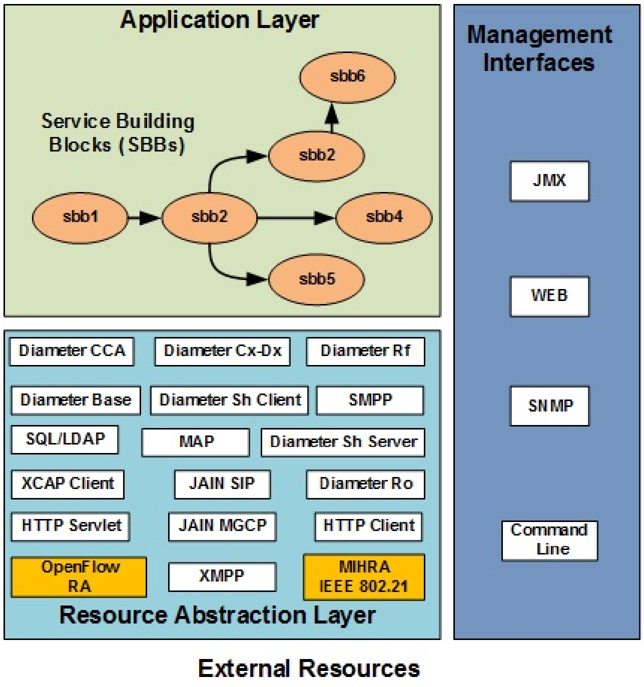


Figure 1: JAIN SLEE Architecture Main Components.

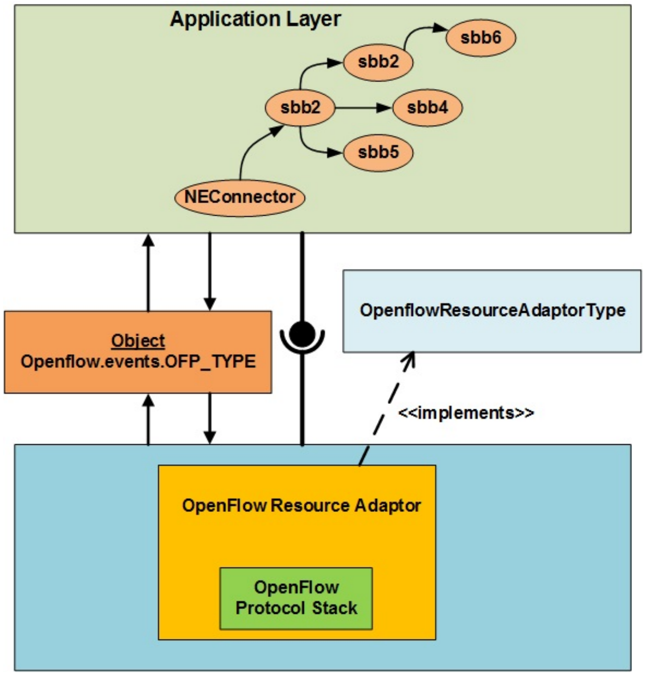


Figure 2: Openflow Resource Adaptor.

such as Rhino [18] and also open source, such as the Mobicents platform [19].

The main components of JAIN SLEE are presented in Fig. 1. The component model consists of two different layers: the application layer that represents the services which run at the JAIN SLEE Application Server and the Resource Adaptation Layer which abstracts underlaying protocol stacks and adapts them to the JAIN SLEE model. The Service Build Block (SBB) contains the application and service logic. Each SBB can be composed of one or more children SBBs and they are organized as a graph. A Service is a deployed and managed artifact which specifies its root SBB and the default event delivery priority. A registered SBB is able to capture and fire events. An Activity is a related stream of events, such as a phone call, that are captured by SBBs entities. The state of these entities can be replicated in a clustered deployment.

According to the JAIN SLEE specification, a *Resource* represents a system that is external to the JAIN SLEE. The Resource Adaptation layer (depicted in Fig. 1) enables several control plane protocols, currently used at the telecommunication protocol stack, to plug in at the JAIN SLEE component model, thus fostering the development of new services and applications that can exploit SDN benefits.

By using clustering, JAIN SLEE supports high availability and fault tolerance. The fault tolerance mode works with state replication and thus a cluster can be viewed as only one virtual container which encompasses all nodes that are active in that cluster. This way, all activity context and SBB entities data are replicated across the cluster nodes. While all the internal components of the JAIN SLEE are fully fault tolerant, the resource adaptors at border with the JAIN SLEE container and the outside environment are not replicated by default, but they can be created to be cluster-aware by using the Fault Tolerant Resource Adaptor API.

Considering all these features, JAIN SLEE is a platform suitable to handle carrier-grade throughput, latency and fault tolerance requirements over a general purpose IT infrastructure [16].

A. *OpenFlow Resource Adaptor*

The OpenFlow protocol and a controller are resources to the JAIN SLEE. To interact with its component model, these resources need to be adapted to its component model, what is accomplished by a *Resource Adaptor* (RA). A RA receives messages from this external system by using a protocol and submits them as events that are produced inside the RA. The RA may, also, consume events created by the services running inside the JAIN SLEE.

The JAIN SLEE specification defines a *Resource Adaptor Type* that basically consists of a set of interfaces that represents common characteristics that must be implemented by a RA of that type. Moreover, the *Resource Adaptor type* references the Events that a *Resource Adaptor* will produce and consume. By using this approach, different resources can be plugged into the JAIN SLEE components.

Usually, resources such as SIP, DIAMETER, XMPP and MGCP protocol stacks have a RA already defined [19]. Regarding OpenFlow there is no *Resource Adaptor* already defined, being this definition one of the contributions of this work. The *OpenFlow Resource Adaptor* (OpenFlowRA), presented in Fig. 2 is responsible for the interaction with the services that run inside JAIN SLEE. The events are captured by the SBB entities according to the service configuration.

The OpenFlowRA implements an *OpenFlowResourceAdaptorType*. This resource type references all the events that might be fired from the OpenFlow stack. In this case, to each message type defined within the *enum ofp_type* at the OpenFlow 1.0 specification [5] an Event that will be sent and also received from the JAIN SLEE was defined.

When the OpenflowRA is actived by the JAIN SLEE, it starts the FloodLight Controller [30]. After this point, all OpenFlow Messages are converted into Events that are sent to the JAIN SLEE. In the same manner the services that run inside JAIN SLEE can dispatch events to the OpenFlowRA that corresponds to OpenFlow messages such as OFPT_FLOW_MOD and OFPT_PACKET_OUT that will be received the OpenFlowRA and then will be forwarded to an OpenFlow Switch by the Controller as OpenFLow Messages.

The implementation uses MessageEvent, a JAIN SLEE feature. Each message received by OpenFlowRA is converted to a JAIN SLEE event. For instance, when a OFPT_PACKET_IN is received by OpenflowRA, it is converted into a MessageEvent of type PacketIn, and it is sent to Abstraction Layer. After be treated by RA, all messages are sent to the Abstraction Layer.

To decouple the services (SBBs) from the signaling control, a default SBB named NEConnector is created. This SBB will be responsible to retrieve the OpenFlow related events, inspect them and fire the corresponding events related to the service that is being created inside a set of SBBs. By using this approach, the NEConnector is the only SBB that needs to care of OpenFlow events. This design enhances further compatibility with new OpenFlow protocol versions, such as OpenFlow 1.3, thus contributing to a low coupling between OpenFlow protocol and other SBBs. The NEConnector abstraction allows the architecture to support several protocols but, if necessary, other SBB can be created to handle other requirements, thus providing a flexible architecture that supports different signaling protocols from the telecommunications world and also from the computer networks world.

IV. Case Study

This case study highlights how the abstraction proposed by this work can be used to integrate OpenFlow with other infrastructure control protocols enabling the deployment of new services and architectures.

The proposed approach was used to integrate OpenFlow with the MIH protocol [20]. The main purpose of the IEEE 802.21 standard for MIH is to facilitate and optimize inter technology handover processes by providing a set of primitives for obtaining link information and controlling link behavior.

The IEEE 802.21 standard offers an abstraction of the control of wireless access links and in this case, an IEEE 802.21 resource adaptor was also created. The extensible component model defined by JAIN SLEE, enabled the integration with this protocol by defining a new resource adaptor. Considering that this RA was not available before, the RA built during this work is also an important contribution to the community that builds JAIN SLEE based services.

The approach envisaged by the OpenFlow was also applied here. Thus the NEConnector is responsible to receive the events generated and route them to the corresponding SBB which is interested in this particular event.

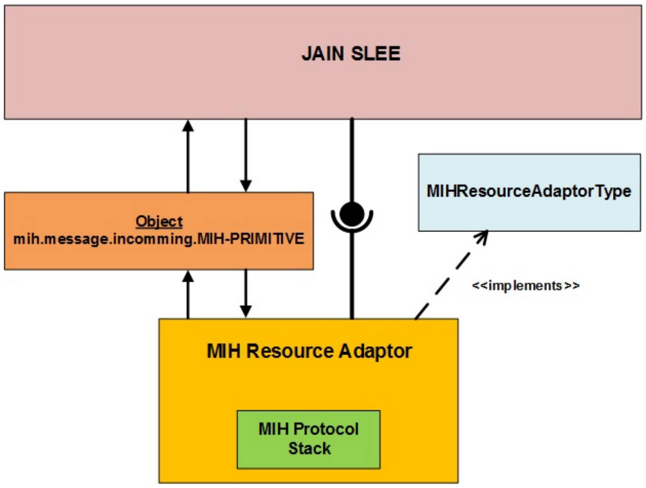


Figure 3: IEEE 802.21 MIH Resource Adaptor.

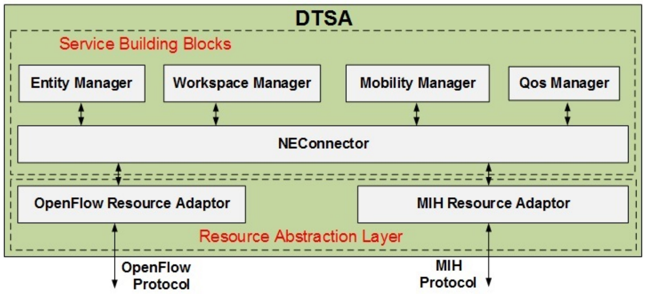


Figure 4: DTSA Components based on JAIN SLEE Component Model.

The MIHRA, as depicted in Fig. 3, implements an *MIHResourceAdptorType*. This resource type references all the events that might be fired from the MIH stack. Thus, there is one different message type to each service primitive defined in the IEEE 802.21 specification.

Moreover, by using the approach proposed in this work and bringing together OpenFlow and also the MIH resource adaptors, it was possible to create the main component of a new network architecture, named Entity Title Architecture (ETArch).

ETArch [45] is a clean slate network architecture, where naming and addressing schemes are based on a topology-independent designation that uniquely identifies an entity. This designation is named Title. ETArch also defines a channel that gathers multiple communication entities. This channel is called Workspace. A key component of this architecture is the Domain Title Service (DTS), which deals with all control aspects of the network. The DTS is composed of Domain Title Service Agents (DTSAs), which maintain information about entities registered in the domain and the workspaces that they are subscribed to, aiming to configure the network devices to implement the workspaces and to allow data to reach every subscribed entity.

The DTSA was created, deployed and tested using the Mobicents JAIN SLEE. Fig. 4 presents the overall DTSA architecture based on the JAIN SLEE component model presented in this work. The NEConnector decouples the DTSA SBBs from the control protocols of the infrastructure, then, the services expressed by the SBBs can also interact with the infrastructure and adapt it, in this case do create the workspace concept on top of a OpenFlow enabled infrastructure and also to support seamless entity mobility by optimizing handover using the IEEE 802.21 MIH protocol. Fig. 4 also presents the SBBs that were created to implement the DTSA, each one dealing with the main aspects of the architecture: the control of entities (EntityManager); the workspaces supported by the DTSA in a given moment (WorkspaceManager); mobility procedures of entities (MobilityManager); and, the QoS enforcement (QoSManager) inside the workspace in order to meet specific communication requirements.

V. Concluding Remarks and Future Work

SDN abstractions can enable new services and applications. Considering OpenFlow, the current materialization of some SDN concepts, the controller is a central piece of the architecture.

The currently available open source controllers are not suitable to meet carrier grade requirements. This kind of controller is a work in progress being conducted by the SDN community through some projects.

This work proposes a new SDN controller architecture, which is integrated with a carrier grade service level execution environment, based on the JAIN SLEE specification. Such controller can abstract several different protocol stacks and provides a common component model where new services and applications could be deployed. JAIN SLEE current implementations are adopted and being used, at plant floor ground, by several telecom operators.

To foster this integration, this work created an OpenFlow resource adapter which enables a cross layer approach where services are able to control the network infrastructure during run-time, by using the OpenFlow protocol.

To showcase the approach presented in this work, an IEEE 802.21 MIH resource adapter was also constructed. The MIH protocol purpose is to abstract the control of wireless access network infrastructure, thus enabling services that need to handle mobility requirements.

By putting together OpenFlow and MIH protocols, the presented case study uses JAIN SLEE SDN capable control layer to deploy a clean slate network architecture named ETArch. In this case, the adopted component model enabled the evolution of the network architecture, enabling new services to be gradually deployed and tested on top of it.

The resource adapters presented here are publicly available, thus collaborating with the research which aims to define, design and deploy next generation computer network architectures.

The carrier grade approach proposed in this work is aligned with most current trends regarding a SDN control layer,but in addition to other proposals, it enables an integration of the protocols that control the network hardware, such as OpenFlow, with the ones the control the applications, thus enabling new types of network services. Moreover, the extensible model can accommodate new protocols and future initiatives, thus preserving investments and becoming an interesting outcome that can be exploited by telecom operators.

As future work, the proposed and constructed carrier grade SDN control layer will be tested under different scenarios in order to demonstrate its fault tolerant and scalability. An OpenFlow 1.3 compliant RA will also be deployed and plugged into the architecture.

The innovative approach proposed under this work presents to the research community a SDN control layer that is suitable to meet carrier grade requirements and is a viable alternative to bring SDN into the telecom infrastructure.

Acknowledgment

This work has been partially funded by the European Community's Seventh Framework Programme, under grant agreement n. 258365 (OFELIA project), by the Brazilian agencies: CAPES, CNPq and FAPEMIG and also by PROPP/UFU.

References

[1] Open Networking Foundation. Software-defined networking: The new norm for networks. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf [retrieved: May, 2014]

[2] G. Goth, "Software-Defined networking could shake up more than packets," IEEE Internet Computing, vol. 15, no. 4, Aug. 2011, pp. 6–9.

[3] M. Palmer. SDN market size to exceed $35b in 2018. [Online]. Available: http://www.sdncentral.com/sdn-blog/sdn-market-sizing/2013/04/ [retrieved: May, 2014]

[4] PLEXXI, LIGHTSPEED, and SDNCentral. SDN market sizing. [Online]. Available: http://cdn.sdncentral.com/wp-content/uploads/2013/04/sdn-market-sizing-report-0413.pdf [retrieved: May, 2014]

[5] N. McKeown et al., "OpenFlow: enabling innovation in campus networks," SIGCOMM Comput. Commun. Rev., vol. 38, no. 2, Mar. 2008, p. 6974, ACM ID: 1355746.

[6] FIBRE. FIBRE Project - Future Internet Testbeds Experimentation Between Brazil and Europe. [Online]. Available: http://www.fibre-ict.eu/ [retrieved: May, 2014]

[7] GENI. OpenFlow - GENI. [Online]. Available: http://groups.geni.net/geni/wiki/OpenFlow [retrieved: May, 2014]

[8] OFELIA. OpenFlow in europe - linking infrastructure and applications. [Online]. Available: http://www.fp7-ofelia.eu/about-ofelia/ [retrieved: May, 2014]

[9] ALIEN. FP7 ALIEN project. [Online]. Available: http://www.fp7-alien.eu/ [retrieved: May, 2014]

[10] S. Sezer et al., "Are we ready for SDN? implementation challenges for software-defined networks," IEEE Communications Magazine, vol. 51, no. 7, Jul. 2013, pp. 36–43.

[11] I. T. UNION, The carrier grade open environment reference model, ser. SERIES Y: Global Information Infrastructure, Internet Protocol Aspects and Next-Generation Networks. International Telecommunication Union, Dec. 2006. [Online]. Available: http://www.itu.int/rec/T-REC-Y.2901-200612-I/en

[12] TeleStax. TeleStax open source cloud communications - success stories. [Online]. Available: http://www.telestax.com/case-studies/ [retrieved: May, 2014]

[13] ORACLE. JAIN general Q&A. [Online]. Available: http://www.oracle.com/technetwork/java/qa-137977.html [retrieved: May, 2014]

[14] D. Ferry. JAIN SLEE (JSLEE) 1.1 specification, final release. [Online]. Available: http://www.jcp.org/en/jsr/detail?id=240 [retrieved: May, 2014]

[15] M. Femminella et al., "Scalability and performance evaluation of a JAIN SLEE-based platform for VoIP services," in Teletraffic Congress, 2009. ITC 21 2009. 21st International, 2009, pp. 1–8.

[16] M. Gomez, E. Torres, J. Chamorro, T. Hernandez, and E. Mendez, "On the integration and convergence of IN and IP mobile service infrastructures," in International Conference on Telecommunications, 2009. ICT '09, May 2009, pp. 143–148.

[17] M. Femminella, R. Francescangeli, E. Maccherani, and L. Monacelli, "Implementation and performance analysis of advanced IT services based on open source JAIN SLEE," in 2011 IEEE 36th Conference on Local Computer Networks (LCN), Oct. 2011, pp. 746–753.

[18] OpenCloud. Rhino SLEE carrier grade system. [Online]. Available: http://www.opencloud.com/products/rhino-application-server/carrier-grade/ [retrieved: May, 2014]

[19] MOBICENTS. Mobicents JAIN SLEE. http://www.mobicents.org/slee/intro.html. [Online]. Available: http://www.mobicents.org/slee/intro.html [retrieved: May, 2014]

[20] IEEE, "IEEE standard for local and metropolitan area networks- part 21: Media independent handover," IEEE Std 802.21-2008, 2009, pp. c1 –301.

[21] N. Gude et al., "NOX: towards an operating system for networks," SIGCOMM Comput. Commun. Rev., vol. 38, no. 3, Jul. 2008, p. 105110. [Online]. Available: http://doi.acm.org/10.1145/1384609.1384625

[22] NOXREPO. About NOX. [Online]. Available: http://www.noxrepo.org/nox/about-nox/ [retrieved: May, 2014]

[23] D. Erickson. Beacon. [Online]. Available: https://openflow.stanford.edu/display/Beacon/Home [retrieved: May, 2014]

[24] Z. Cai. Maestro. [Online]. Available: http://code.google.com/p/maestro-platform/ [retrieved: May, 2014]

[25] NTT Communications. Ryu SDN framework. [Online]. Available: http://osrg.github.io/ryu/ [retrieved: May, 2014]

[26] H. Shimonishi, Y. Chiba, Y. Takamiya, and K. Sugyo. Trema repository. [Online]. Available: http://trema.github.io/trema/ [retrieved: May, 2014]

[27] Travelping. FlowER - erlang OpenFlow development platform. [Online]. Available: http://travelping.github.io/flower/ [retrieved: May, 2014]

[28] D. Staessens, S. Sharma, D. Colle, M. Pickavet, and P. Demeester, "Software defined networking: Meeting carrier grade requirements," in 2011 18th IEEE Workshop on Local Metropolitan Area Networks (LANMAN), 2011, pp. 1–6.

[29] F. Tam, "On engineering standards based carrier grade platforms," in Proceedings of the 2007 workshop on Engineering fault tolerant systems, ser. EFTS '07. New York, NY, USA: ACM, 2007. [Online]. Available: http://doi.acm.org/10.1145/1316550.1316554

[30] Big Switch Networks. Floodlight OpenFlow controller. http://floodlight.openflowhub.org/. [Online]. Available: http://floodlight.openflowhub.org/ [retrieved: May, 2014]

[31] T. Koponen et al., "Onix: a distributed control platform for large-scale production networks," in Proceedings of the 9th USENIX conference on Operating systems design and implementation, ser. OSDI'10. Berkeley, CA, USA: USENIX Association, 2010, p. 16. [Online]. Available: http://dl.acm.org/citation.cfm?id=1924943.1924968

[32] OpenDaylight. OpenDaylight technical overview. [Online]. Available: http://www.opendaylight.org/project/technical-overview [retrieved: May, 2014]

[33] ON.LAB. ONOS - open network operating system. [Online]. Available: http://tools.onlab.us/onos.html [retrieved: May, 2014]

[34] B. S. Networks. Big network controller. [Online]. Available: http://www.bigswitch.com/products/SDN-Controller [retrieved: May, 2014]

[35] NICIRA. Nicira. [Online]. Available: http://nicira.com/ [retrieved: May, 2014]

[36] O. Research. Modern SDN stack project. [Online]. Available: http://onrc.stanford.edu/research_modern_sdn_stack.html [retrieved: May, 2014]

[37] C. Matsumoto. What OpenDaylight really wants to do. [Online]. Available: http://www.lightreading.com/blog/software-defined-networking/what-opendaylight-really-wants-to-do/240152993 [retrieved: Apr., 2014]

[38] ON.LAB. ONOS at ONS 2014. [Online]. Available: http://www.slideshare.net/ON_LAB/onos-at-ons-2014 [retrieved: Mar., 2014]

[39] E. Schooler et al. SIP: session initiation protocol. [Online]. Available: http://tools.ietf.org/html/rfc3261 [retrieved: May, 2014]

[40] J. Arkko, E. Guttman, P. R. Calhoun, and J. Loughney. Diameter base protocol. [Online]. Available: http://tools.ietf.org/html/rfc3588 [retrieved: May, 2014]

[41] P. Saint-Andre. Extensible messaging and presence protocol (XMPP): core. [Online]. Available: http://tools.ietf.org/html/rfc6120 [retrieved: May, 2014]

[42] B. Foster and F. Andreasen. Media gateway control protocol (MGCP) version 1.0. [Online]. Available: http://tools.ietf.org/html/rfc3435 [retrieved: May, 2014]

[43] F. de Oliveira Silva, J. de Souza Pereira, P. Rosa, and S. Kofuji, "Enabling future internet architecture research and experimentation by using software defined networking," in 2012 European Workshop on Software Defined Networking (EWSDN), 2012, pp. 73–78.

[44] ORACLE. JAIN API specifications. [Online]. Available: http://www.oracle.com/technetwork/java/api-specs-137688.html [retrieved: May, 2014]

[45] C. Guimarães et al., "IEEE 802.21-enabled entity title architecture for handover optimization," in IEEE WCNC'14 Track 3 (Mobile and Wireless Networks) (IEEE WCNC'14 Track 3 : NET), Istanbul, Turkey, Apr. 2014, unpublished article. [Online]. Available: http://www.facom.ufu.br/~flavio/wcnc-2014/IEEE_802.21-enabled_ETArch_for_Handover_Optimization.pdf

A Bluetooth Network Dynamic Graph

Celio Marcio Soares Ferreira,
Ricardo Augusto R. Oliveira,
Haroldo Santos Gambini
Computer Science Department
Federal University of Ouro Preto (UFOP)
Ouro Preto, Minas Gerais, Brasil
e-mail: celio@linuxplace.com.br,
{rrabelo, haroldo.santos}@gmail.com

Alejandro C. Frery
Instituto de Computação
Universidade Federal de Alagoas (UFAL)
Maceió, AL, Brasil
e-mail: acfrery@gmail.com

Saul Delabrida,
Mateus Freire Carneiro
Computer Science Department
Federal University of Ouro Preto (UFOP)
Ouro Preto, Minas Gerais, Brasil
e-mail: saul@sdelabrida.com,
mateusfreire05@hotmail.com

***Abstract*—Bluetooth uses a communication technique called frequency-hopping which has some collateral effects. One of these is a significant delay time during the phase of discovery of nodes. To precisely estimate this delay, we use real and simulated devices, and we measure the elapsed time until a Piconet formation. Our contribution is modeling the Bluetooth network as a dynamic graph, adding the frequency-hopping procedures, Piconet limits and network constraints. These new components render a model which is more consistent with the Bluetooth network technology specification than those presented so far. Our graph can be used as a basis for realistic optimization models.**

***Keywords*—bluetooth; scatternet; frequency hopping; dynamic graph.**

I. Introduction

In 2012, 1.1 billion mobile phones were shipped, almost 100 percent with Bluetooth technology [1]; but, despite this popularity, Bluetooth network applications are not yet explored in their full potential. The new-coming wearable devices, like smart-watches, and smart-glasses to name a few, use Bluetooth intensively. New popular apps, like firechat, also depend on the ad hoc Bluetooth network formation. The possibility of forming wider-ranging ad hoc networks enhances their most common use: files exchange apps and mono headsets.

During communication, Bluetooth devices do not use a fixed frequency; they use frequency-hopping. Therefore, for a link formation, a device discovery phase is firstly needed.

In the device discovery phase, one device scans for another device, and both send and listen messages, respectively, in a pseudo-random frequency sequence, until a frequency coincidence occurs and the synchronization messages are delivered. Even in a smallest network with two nodes, there is a delay time as consequence of the randomness, this prohibits Bluetooth for latency-sensitive applications.

The frequency-hopping sequence and all data flows are coordinated centrally by a node called master, in a master-slave point-to-multipoint network called Piconet. Each Piconet contains only one master, and can have a maximum of seven active slave nodes communicating to $10\,\mathrm{m}$ range.

To expand the limits of this communication, we prefer Scatternets. They are collections of Piconets joined by a Bridge node and coordinated by a protocol.

The collateral effects of frequency-hopping in Bluetooth, such as the delay in discovery phase, show the relevance of a Bluetooth network graph model. It is important to devise a realistic model, having procedures and constraints consistent with the technology specification.

In this work, we have the following contributions:

A Bluetooth network dynamic graph, with:

- The Piconet and Scatternet characterization and topology constraints;
- Master and slaves node rules;
- The proposal, implementation and validation of two procedures: $\mathrm{FHS}()$ and $\mathrm{Disc}()$, which represent a device frequency hopping sequence synchronization and discovery of devices.

These new procedures and network constraints characterize our Bluetooth network graph as a dynamic graph.

In Section II, we describe the initial formation process of a Piconet. Section III discusses the related work. In Section IV the initial connection delay time is identified by simulation and real experiments. In Section V, we present the dynamic Bluetooth network graph model. Finally, the conclusions are in Section VI.

II. Piconet - the basic Bluetooth Network formation

In a Piconet, a device assumes the role of master or slave, and two distinct phases are required to connect: the discovery and the link formation.

During the discovery phase, the candidate for the master device goes into the **INQUIRY** state, looking for the devices candidates for the slave in an **INQUIRY SCAN** state.

During the **INQUIRY** state, the searching device sends an IDentifier (ID) by broadcast using 32 of the 79 frequencies defined by Bluetooth specification. The sequence of frequencies that will be used to broadcast ID messages, is a pseudo-random calculate, derived from the clock of the device. The set of this frequencies is called Inquiry Hopping Sequence (IHS).

The candidate for slave device in **INQUIRY SCAN** state listens to broadcasts ID, on the same 32 IHS frequencies, hopping in a pseudo-random sequence derived from its clock.

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-360-5

In this phase, candidates for master and slaves send and listen messages respectively in a sequence of pseudo-random frequency hops, until a frequency coincidence occurs. A time slot difference collaborates with the increased likelihood of the device hearing the same channel on which a ID was listened: the devices in **INQUIRY** state hop in time slots of $312.5\mu s$ faster than the standard Bluetooth $625\mu s$ used by devices in **INQUIRY SCAN**.

After receiving an ID, the candidate for slave device assumes a state called **INQUIRY RESPONSE**, and responds to the request by sending its network address and clock, in a packet called frequency-hopping synchronization (FHS), using the same frequencies of IHS. Then, the device waits for the backoff a random value of time slots $(0 - 639.375)\mu s$, with the objective of minimizing collisions of responses, and goes to **PAGE SCAN** state.

When the candidate for master device receives the FHS, it enters a state of **PAGE**, and uses the information received from the FHS for synchronization and connection with the candidates for slave device that have already been discovered and are in the **PAGE SCAN** state.

During the **PAGE** state, the candidate for master device selects a candidate for slave device to be connected sending packages to the candidate for slave devices previously discovered using the sequence of estimated hops.

After the **PAGE** process is complete, the Piconet is formed and the devices gain a connected status and assume their master and slaves roles.

III. Related Work

Jedda, Jourdan and Zaguia [2] analysed the impacts of changing Bluetooth parameters on the static and dynamic Scatternet formation protocols. These parameters are related to the use of the frequency hop communication technique. The Scatternet formation on static protocols happens as follows; each node alternates randomly between the **INQUIRY** and **INQUIRY SCAN** Bluetooth discovery states, when one device discovers each other, a temporally Piconet is formed until being destroyed at the end of the communication. They called this mechanism of *ALTERNATE*, being examples of it: BlueStars Petrioli, Basagni and Chlamtac [3]; BlueMIS Zaguia, Stojmenovic and Daadaa [4] and BlueNet Wang, Thomas and Haas [5]. In dynamic Scatternet protocols, the discovery phase is interlaced with the network formation, the node shares its time between discovering new devices and communication in the Scatternet. The examples of dynamic protocols are: Law, Mehta and Siu [6] and Cuomo, Melodia and Akyildiz [7]. Jedda, Jourdan and Zaguia [2] using ns-2 [8] simulator, found that changing parameters of Bluetooth 1.2 discovery phase, produce *ALTERNATE* Scatternets 3.5 times faster.

In Pettarin, Pietracaprina and Pucci [9], the Bluetooth dynamic topology was described as a sequence of graphs $\mathcal{G}(n, \rho, r(n), c(n), \epsilon) = \{G_t : t \in \mathbb{N}\}$. Each of the n agents moves to a grid node chosen uniformly at random among the grid nodes within euclidean distance ρ from its current position, being t each time step, linked to the movement of devices. This model describes situations in which the devices are moving and establishing network connections. This sequence of graphs can be modeled as a Markov chain Clementi, Monti, Pasquale e Silvestri [10], whose transitions describe by the model of moving nodes.

Ferraguto, Mambrini, Panconesi and Petrioli [11] proposes a Bluetooth network graph model, where the links can be described by n devices randomly distributed in the unit square, the function $c(n)$ is the neighborhood of each device and $r(n)$ is the range of each device. With this, the Bluetooth network is denoted by the graph $BT(r(n), c(n))$.

Jedda, Jourdan and Zaguia [2] show, by ns-2 simulation, that the parameters changes related to Bluetooth discovery phase, are more significant in static Scatternet protocols, that use use the ALTERNATE strategy, than in dynamic Scatternet protocols as in Law, Mehta and Siu [6]. This opens a discussion about the importance of new propositions that include changes in parameters related to the Bluetooth frequency-hopping.

The Random Geometric Graph (RGG) has been employed for the characterization of Bluetooth network topology. Pettarin, Pietracaprina and Pucci [9] discuss the expansion and diameter of RGG subgraphs induced by device discovery phase. Experimental evidence shows that $BT(r(n), c(n))$ is a ideal model for the Bluetooth topology; see Ferraguto, Mambrini, Panconesi and Petrioli [11]. Unlike our proposed model, classical graph models of Bluetooth as Gupta and Kumar [12], Ferraguto, Mambrini, Panconesi and Petrioli [11], Crescenzi, Nocentini and Pietracaprina [13] and Pettarin, Pietracaprina and Pucci [9], do not explore (i) the topology of a Piconet, (ii) the Scatternet formation, or (iii) the intrinsic influences of frequency-hopping communication.

A correct mapping of frequency-hopping peculiarities is essential for the suitable design of Bluetooth solutions and applications.

IV. Identifying the delay

In order to measure the elapsed time of a Piconet connection and verify the existence of a connection delay, we generated Piconet instances, using simulation and real devices.

We used the UCBT [14], a ns-2 [8] extension that simulates Bluetooth, developed by the University of Cincinnati. Additionally, to verify the real scenario, we used the Lego NXT Mindstorm [15], and robots were assembled to establish connections with each other via Bluetooth. The Bluetooth communication interface in Lego Mindstorms NXT kits has the channels according with the Bluetooth specification, however the buffer of the equipment does not allow the use of more than three channels simultaneously. Piconets were testes limited to three devices. The robots were assembled as vehicles, and the Bluetooth configured for creating the the communication link as soon as they entered the network range. Once the connection is established, the vehicles should move in opposite directions until losing the communication due to distance. Once the connection between the robots stops, they must return to the starting point to restore the Piconet. Connections were restored even in the presence of thick walls. All the elapsed times were collected while the robots were moving.

In the first experiment, we generated 30 real and simulated instances of Piconets with sizes of two (one master and one slave) and three nodes (one master and two slaves). We measured the amount of time until all the slave nodes were in connected status with the master node. The box-plots in Figure 1 and Figure 2 represent the elapsed time until total connection of nodes in the 30 instances of each Piconet size. We observe that the elapsed time until total Piconet connection is, in mean, one second longer, even between one master-slave link. This is a problem because long connection time is not tolerated by many security and medical applications, among others.

In a second experiment, we generated 30 ns-2 simulation instances to each Piconet formed with one master and n neighboring candidates for the role of slave device. In accordance with Pettarin, Pietracaprina and Pucci [9], we observed that as the value of n increases, so is the likelihood of connection. The rationale for this is that during the discovery phase, all devices in **INQUIRY SCAN** perform pseudo-random hops in slower time slots than the master until there is a match of frequencies. This behavior shows that, despite the increase in density of devices within the range of the master, the frequency-hopping technique provides greater resilience to collisions and depletion of the spectrum. Figure 3 show the elapsed time until first **INQUIRY RESPONSE**.

Figure 4 shows the formation of a theoretically maximal Piconet, represented by one master and seven slaves. We observed the proportional increase value of n and the time. This behaviour is explained by the need for matching the channel in the discovery phase, the backoff and a scheduling of intra-Piconet synchronization packets.

In order to create a new slave entry in an existing Piconet, the master needs to stop the intra-Piconet communication and a new discovery must start. The slaves that have already entered the Piconet change to **HOLD** mode, waiting for new polling from the master before re-communicating. The time cost of this operation grows with the increase of devices due to the new discovery and resynchronization by the intra-Piconet Scheduling process.

The error bars in Figure 4 show the high degree of variability and delay in the connection, represented by random variables associated with the discovery of slaves, backoff time and intra-Piconet scheduling processes.

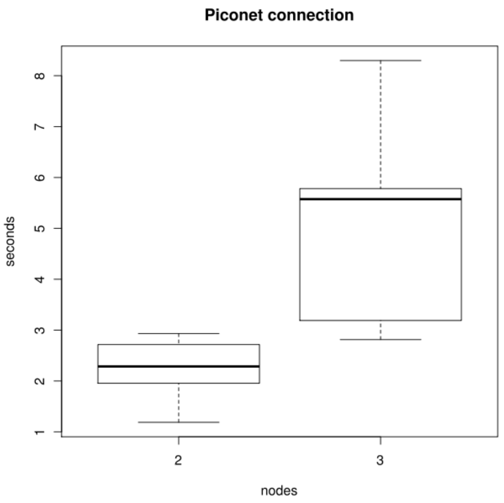


Figure 1. Elapsed Time until complete Piconet connection using Lego NXT.

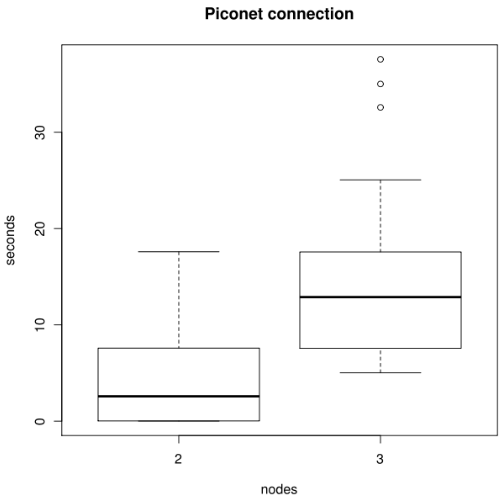


Figure 2. Elapsed Time until complete Piconet connection using ns-2

V. Dynamic Graph of Bluetooth network

The Bluetooth network will be described as a graph, following the definition by Gupta and Kumar [12] and Pettarin, Pietracaprina and Pucci [9].

Consider the undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}'\}$ composed by the set of $n \geq 1$ nodes $\mathcal{V} = \{v_1, \dots, v_n\}$ and of edges $\mathcal{E}' \subset \{\mathcal{V} \times \mathcal{V}\}$, such that $(e_i, e_j) \in \mathcal{E}' \iff (e_j, e_i) \in \mathcal{E}'$. A dynamic Bluetooth graph, able to describe the formation of Piconets and Scatternets, will be defined on top of this generic graph with the inclusion of a spatial restriction and of nodes labels.

A common assumption for the placement of nodes is the fully independent or binomial model Ramos, Guidoni, Nakamura, Boukerche and Frery [16]. According to this model, given n nodes, their coordinates $(x_i, y_i)_{1 \leq i \leq n}$ in the $[0,1]^2$ square are outcomes of $2n$ independent identically distributed uniform $[0,1]$ random variables. Nodes represent devices, and once they are deployed, the links which enable the communication are built according to the range $r(i)$ of each device. The geometric rule says that the edge $(e_i, e_j) \in \mathcal{E}'$ may may exist only if $d(v_i, v_j) \leq \min\{r(i), r(j)\}$, i.e., nodes v_i and v_j may communicate only if both can talk to each other. The distance function $d\colon [0,1]^2 \times [0,1]^2 \to \mathbb{R}_+$ is arbitrary and may incorporate any prior information about the environment as, for instance, obstacles. The choice of the unitary square support does not impair any lack of generality on the model. Many applications assume reciprocal communication setting $r(i) = r$ for every $1 \leq i \leq n$.

The definition of protocols for the operation of Bluetooth networks requires another ingredient: identifying masters and slaves. Each node receives a label, either "M" or "S" for denoting its current state.

A new graph can be now defined, provided the graph $\mathcal{G}$ defined as above. The Bluetooth graph BT is a subset of $\mathcal{G}$

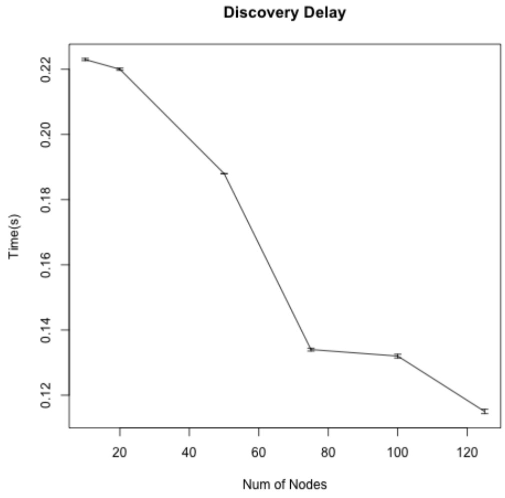


Figure 3. Time until first **INQUIRY RESPONSE**

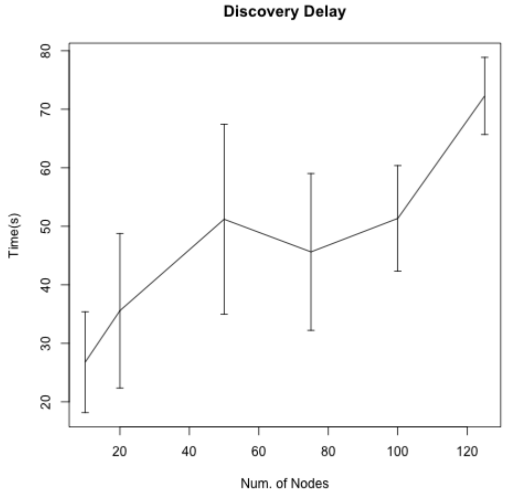


Figure 4. Time until formation of the first complete Piconet with 7 slaves and 1 master

such that $BT = \{\mathcal{V}, \mathcal{E}\}$ since possibly not all allowed connections are set, i.e., $\mathcal{E} \subset \mathcal{E}'$. The communication specification $\mathcal{E}$ has to satisfy the following requirements:

1) There is at least one M node.
2) All S nodes connect to one and only one M node.
3) M nodes do not have connections among them.
4) S nodes do not have connections among them.

A Piconet if formed if there is only one M node and all S nodes connect to it. If there is more than one M node, we are in the presence of a Scatternet.

Pettarin, Pietracaprina and Pucci [9] describe situations in which the devices are moving. The BT graph is, thus, dynamic and can be described as a function of the time $BT(t) = \{\mathcal{V}, \mathcal{E}(t)\}$. The authors describe the sequence of graphs by means of a Markov chain whose transitions express the change of connections due to the movement of the nodes.

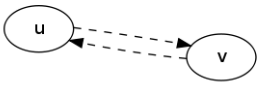


Figure 5. The nodes u and v are in range of each other

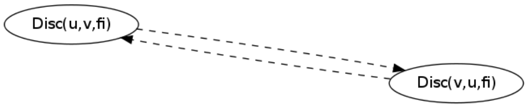


Figure 6. The nodes u and v initialize the discovery $\mathrm{Disc}()$

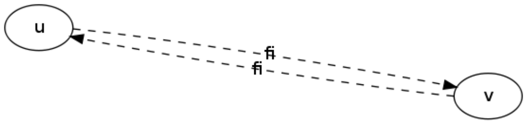


Figure 7. The coincidence of frequency f_i occurs

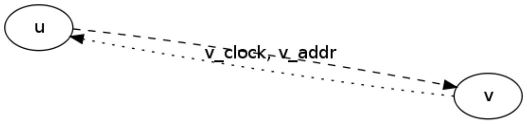


Figure 8. The v node transmit its network address and clock

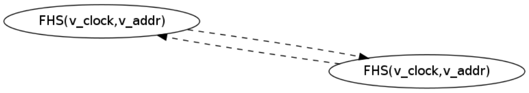


Figure 9. The nodes use FHS() to generate the F' frequency sequence

Denote $F = \{f_i : 0 \leq i \leq 79\}$ the set of frequencies used in FHSS, then FHS is a function $FHS(CLK, MS)$, where CLK is the clock of the elements involved and MS is the address of the Piconet master. The details of FHS are given by the Bluetooth specification. Each master-slave link has an unique pseudo-random sequence of frequencies $F' = (f'_1, f'_2, \dots)$, so that $f'_i \in F$.

Let u and v to nodes, as in Figure 5. We define the process of discovery as the operation $\mathrm{Disc}(u, v, f_i)$ which consists of the insertion of the edge (u, v) in $\mathcal{E}$.

The $\mathrm{Disc}()$ process has its execution distributed, while running u and v at the same time, see Figure 6. Master and slaves begin a sequence of pseudo-random frequency hops, until a frequency f_i coincidence occurs, as illustrated in Figure 7. After matching, the slave waits for a random time to respond $FHS()$ to the master, this is called Backoff interval. This is necessary because the $FHS()$ must be exchanged between nodes; see Figures 8 and 9. For this purpose, after $\mathrm{Disc}()$ has been applied, the slave returns $FHS()$ and generates the correct pseudo-random sequence F' for the connection; see Figure 10. The labels “M” and “S” are given to the nodes selected as master and slave, respectively, see Figure 11.

Changes as the ones described above in the connectivity of the Bluetooth network make it a dynamic graph; see Frigioni and Italiano [17].

Figure 10. Now the nodes use the frequency sequence F' to exchange messages

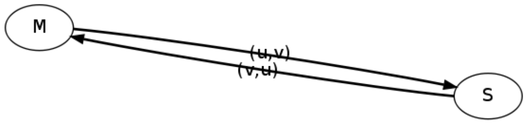


Figure 11. After frequency synchronization, the nodes receive the labels M-master and S-slave, and the links represented by the edges (u, v) and (v, u) begin the message transport

VI. Conclusion and Future Work

The delay observed during the initial Bluetooth connection process is directly related to the effects of frequency-hopping technology use. This delay is the sum of:

1) in Bluetooth discovery phase:
   - a random value of time until the coincidence of frequencies between listener and sender devices;
   - the random value of time of Backoff until node can listen a response;
2) in Piconet with more than two nodes, during the entry of a new slave in a existing Piconet, a new discovery process is needed, and a intra-Piconet scheduling for master frequency resynchronization.

The randomness in discovery and its collateral effect, the delay, is a crucial constraint for simple Piconet applications that require adequate responsiveness. The $\mathrm{Disc}()$ and $\mathrm{FHS}()$ functions in the dynamic graph, shown in Section V, are the procedures affected.

The problem of delay and specific characteristics of a network using frequency-hopping as Bluetooth Piconet, shows the relevance of a model consistent with the Bluetooth specification as our dynamic graph.

New research proposing changes to the Bluetooth specification need to be leveraged. The search for techniques that reduce the duration of discovery phase, will give rise to new use cases for Bluetooth network.

As future works, we will add the **PAGE** procedures to Bluetooh network dynamic graph, and use it as the basis of the constraints in the Ferreira, Oliveira, Gambini and Frery [18].

Acknowledgement

Thanks to Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), to CNPq, CAPES, UFOP (Universidade Federal de Ouro Preto), SEVA Engenharia and FAPEAL.

References

[1] bluetooth.com, ”The bluetooth network effect,” Last Visited in 05/30/2014. [Online]. Available: http://www.bluetooth.com/Pages/network-effect.aspx/

[2] A. Jedda, G.-V. Jourdan, and N. Zaguia, ”Some side effects of fhss on bluetooth networks distributed algorithms,” in Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications - AICCSA 2010, ser. AICCSA 10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 1–8.

[3] C. Petrioli, S. Basagni, and I. Chlamtac, Configuring bluestars: multihop scatternet formation for bluetooth networks, Computers, IEEE Transactions on, vol. 52, no. 6, 2003, pp. 779-790.

[4] N. Zaguia, I. Stojmenovic, and Y. Daadaa, Simplified bluetooth scatternet formation using maximal independent sets, in Complex, Intelligent and Software Intensive Systems, 2008. CISIS 2008. International Conference on, 2008, pp. 443–448.

[5] Wang, Zhifang and Thomas, Robert J. and Haas, Zygmunt J., ”Bluenet - A New Scatternet Formation Scheme,” in HICSS , 2002, pp. 9 pp.

[6] C. Law, A. K. Mehta, and K.-Y. Siu, ”A new bluetooth scatternet formation protocol,” Mob. Netw. Appl., vol. 8, no. 5, Oct. 2003, pp. 485–498,

[7] F. Cuomo, T. Melodia, and I. Akyildiz, Distributed self-healing and variable topology optimization algorithms for qos provisioning in scatternets, Selected Areas in Communications, IEEE Journal on, vol. 22, no. 7, 2004, pp. 1220-1236.

[8] N. S. 2, ”The network simulator - ns2,” Last Visited in 05/30/2014. [Online]. Available: http://www.isi.edu/nsnam/ns/

[9] A. Pettarin, A. Pietracaprina, and G. Pucci, ”On the expansion and diameter of bluetooth-like topologies,” in Algorithms - ESA 2009, ser. Lecture Notes in Computer Science, A. Fiat and P. Sanders, Eds. Springer Berlin / Heidelberg, 2009, vol. 57, pp. 528–539.

[10] A. Clementi, A. Monti, F. Pasquale, and R. Silvestri, Information spreading in stationary markovian evolving graphs, Parallel and Distributed Systems, IEEE Transactions on, vol. 22, no. 9, 2011, pp. 1425-1432.

[11] F. Ferraguto, G. Mambrini, A. Panconesi, and C. Petrioli, A new approach to device discovery and scatternet formation in bluetooth networks. in IPDPS. IEEE Computer Society, 2004.

[12] P. Gupta and P. Kumar, ”The capacity of wireless networks,” Information Theory, IEEE Transactions on, vol. 46, no. 2, Mar. 2000, pp. 388–404.

[13] P. Crescenzi, C. Nocentini, A. Pietracaprina, and G. Pucci, On the connectivity of bluetooth-based ad hoc networks, Concurrency and Computation: Practice and Experience, vol. 21, no. 7, 2009, pp. 875-887.

[14] D. A. Q. Wang, ”Ucbt - bluetooth extension for ns2 at the university of cincinnati,” Last Visited in 05/30/2014. [Online]. Available: http://www.cs.uc.edu/cdmc/ucbt/

[15] lego.com, ”Lego - NXT - Software,” Last Visited in 05/30/2014. [Online]. Available: http://www.lego.com/en-us/mindstorms/downloads/nxt/nxt-software/

[16] H. S. Ramos, D. L. Guidoni, E. F. Nakamura, A. Boukerche, A. C. Frery, and A. A. F. Loureiro, Topology-related modeling and characterization of wireless sensor networks, in PE-WASUN2011 ACM International Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks. Miami, EUA: 2011ACMi, 2011.

[17] D. Frigioni and G. F. Italiano, ”Dynamically switching vertices in planar graphs (extended abstract),” in Proceedings of the 5th Annual European Symposium on Algorithms. London, UK: Springer-Verlag, 1997, pp. 186–199.

[18] C. M. Soares Ferreira, R. A. Rabelo Oliveira, H. S. Gambini, and A. C. Frery, Static bluetooth scatternet formation models: The impact of fhss, in AICT 2013, The Ninth Advanced International Conference on Telecommunications, 2013, pp. 25-31.

Energy-free Security in Wireless Sensor Networks

Adel Elgaber
Institut FEMTO-ST UMR CNRS 6174
Université de Franche-Comté, France
aelgaber@femto-st.fr

Julien Bernard
Institut FEMTO-ST UMR CNRS 6174
Université de Franche-Comté, France
julien.bernard@femto-st.fr

Yacouba Ouattara
Institut FEMTO-ST UMR CNRS 6174
Université de Franche-Comté, France
youattar@femto-st.fr

***Abstract*—Wireless sensor networks are often deployed in open and uncontrolled environments that make them more vulnerable to security attacks. Cryptographic algorithms can be used to protect the data collected by the sensors against an intruder. The cost in terms of energy to provide enough security can be quite large as these algorithms may be very complex. As communication is the main energy consumer, a way to save energy is to use data compression. We propose to measure the impact of the well-known DES algorithm on the energy consumption for various number of rounds and then, we show that energy-free security may be possible. We combine a cryptographic algorithm with a compression algorithm and show through a model that a node can provide security without consuming more energy. The only counterpart is the time for ciphering and compressing. We get some results from experiments on energy consumption of cryptographic and compression algorithms and establish the level of security that can be achieved in various cases, from a single node to a random network.**

***Keywords*–Wireless sensor networks; security; compression; energy**

I. Introduction

A wireless sensor network (WSN) is a specific ad-hoc network with a large number of nodes that have limited energy. These networks are used for collecting information about natural phenomena and other applications. As they are generally deployed in open and uncontrolled environments, wireless sensor networks are more vulnerable to security attacks [1][2][3][4].

Many cryptographic protocols have been proposed to deal with security issues [2][5][6]. They rely on cryptographic primitives like public key cryptographic algorithms or secret key cryptographic algorithms or cryptographic hash functions. The impact on energy consumption of some of these primitives has already been evaluated [7].

As communication is the main energy consumer, a way to save energy in wireless sensor network is to use data compression [8]. Again, the impact on energy consumption of compression algorithms has been evaluated [9][10][11].

In this paper, we first propose to measure the impact of the well-known Data Encryption Standard (DES) algorithm on the energy consumption of a MSP430-based node. Then, our main contribution is to show that energy-free security is possible. We combine a cryptographic algorithm with a compression algorithm and show through a model that a node can provide security without consuming more energy. The only counterpart is the time for ciphering and compressing involving CPU cycles, which is largely less consuming than communication.

In section II, we analyze the related work regarding security and compression algorithms in wireless sensor networks. Then, in section III, we provide experimental results for DES on a MSP430 based node from the Senslab plaform. In section IV, we give an energy consumption model for cryptographic algorithms and compression algorithms and in section V, we show that it may be possible to have energy-free security. In section VI, we use a linear network to achieve better security considering the energy of the whole network. Finally, in section VII, we show that, even on random networks, it is possible to have strong energy-free security with high probability.

II. Related work

Regarding security in general, there are two main families of cryptographic algorithms: public key cryptographic algorithms like RSA or ElGamal and secret key cryptographic algorithms like DES or Advanced Encryption Standard (AES). The advantages of public key cryptography is the availability of authentication and key exchange mechanisms. Public key cryptography is secure and reliable as it is based on strong mathematical theorems. Meanwhile it needs complex arithmetical and logical operations. Strong public key cryptography can affect the lifetime of a node [12][4].

The other solution is to use secret key cryptography. Secret key cryptography relies on simple operations like bit shifting and basic bitwise logical operations (or, and, xor) that can easily be adapted to sensor nodes. Lee et al [7] considered some well-known cryptographic algorithms (AES, RC5, Skipjack, XXTEA) and studied the influence of some parameters (number of rounds, size of the key) on the energy consumption of MicaZ and TelosB sensor nodes. In particular, they show that the energy consumption of RC5 encryption increases linearly with the number of rounds.

In a wireless sensor node, communication is the main energy consumer. An idea to save energy is to apply a compression algorithm before sending data so that the energy used for compression is counterbalanced by the energy saved for communication [8][11]. Capo et al. [9] used a MSP430-based node and measured the consumption of different compression algorithms: S-LZW, Run-Length Encoding (RLE) and K-RLE.

III. Experiments with DES on MSP430

A. Data Encryption Standard (DES)

DES is a symmetric key cryptographic algorithm that was standardized in 1977 and has been used widely since then. DES is based on a Feistel scheme with a 56-bit key and 16

TABLE I. ENERGY CONSUMPTION OF DES WITH VARIOUS NUMBER OF ROUNDS

Rounds	2	4	8	16
Energy (μJ)	21.48	24.75	30.27	40.55

rounds. Each round consists in a fixed set of four operations called Expansion, Key mixing, Substitution and Permutation that operates on 64-bit blocks that are divided in two 32-bit half-blocks [13]. In our experiments, we use a custom implementation of DES in C that can be tuned to reduce the number of rounds.

B. The Senslab testbed

The Senslab platform [14] is an experimental platform for wireless sensor networks. Its aim is to automate the deployment, test, and monitoring of wireless sensor network applications. Each of the four sites of the platform has 256 MSP430-based nodes that can be used to make tests. Each node can be monitored with several probes. The frequency of the measures can be chosen for each experiment. At the end of the experiment, the measures are stored in a simple file for each node.

In our experiment, we used the Senslab platform and we measured the power consumption of a node that was compressing some data with the DES algorithm. The frequency of the measures was set to 100ms.

C. Experiment description

The experiment consists in measuring the power consumed by the DES algorithm on a MSP430-based node of the Senslab platform. For each experiment, we used random input data of $\lambda = 64$ bits. The number of rounds ranges over the values 2, 4, 8 and 16. Each experiment is repeated five times and the average value is given.

D. Results

Table I shows the results obtained in the previous experiment with various number of rounds.

Figure 1 shows the same results on a plot. We observe that the relation between energy consumption and the number of rounds is linear, of the form: $E = E_0 + r \times E_r$ where r is the number of rounds. E_0 is a constant that represents the energy for constant operations in the algorithm, i.e., operations that do not depend on the number of rounds : initial and final permutation, permuted choice 1 (PC1) in the key schedule. E_r is the energy per round. Applying a linear regression on the data of table I, we find $E_0 = 19.149\mu$J and $E_r = 1.348\mu$J/round. The estimation of this linear regression is also shown on figure 1.

As a conclusion of this experiment, we found a linear relationship between energy consumption and the number of rounds for DES. It is of the same kind as the one found in [7] for RC5. These uncorrelated results can lead to a general model for encryption with symmetric cryptographic algorithm. This model is described in the next section.

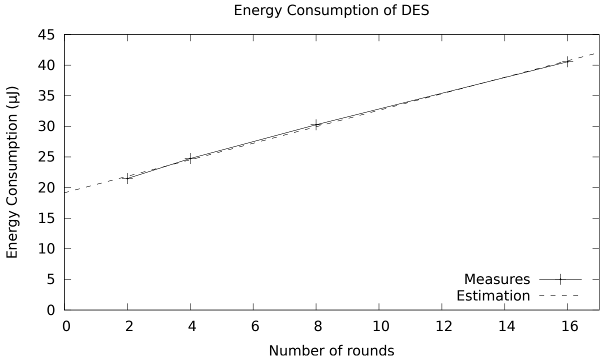


Figure 1. Energy consumption of DES with various number of rounds

TABLE II. VALUES OF E_0^{ENC} AND E_r^{ENC} FOR VARIOUS ENCRYPTION ALGORITHMS

Algorithm	E_0^{enc} (nJ/bit)	E_r^{enc} (nJ/round/bit)
DES	299.2	21.06
RC5 [7]	336.4	173.28

IV. MODELING COMPRESSION AND ENCRYPTION IN WSN

A. Modeling encryption

As seen before, we can model the energy consumption of encryption as:

$$E^{\text{enc}}(\lambda, r) = \lambda \times (E_0^{\text{enc}} + r \times E_r^{\text{enc}}) \tag{1}$$

In addition to the previous experiment, we introduce λ, the length of the input data, in our model. There should be another constant factor that does not depend on the length of the input data, but we assume this constant factor is negligible. For example, in DES, the only operation that does not depend on the number of rounds and the length of the input data is PC1, which is a very simple operation compared to the rest of the algorithm.

Table II shows the values of E_0^{enc} and E_r^{enc} for various algorithms. The DES values are computed from our experiment. The RC5 values are computed from figure 3 in [7]. We took the values of the TelosB node as it is a MSP430-based node, and we summed the ”setup” phase and ”encryption” phase to have the full algorithm. We assumed the length of data to be $\lambda = 8$ bytes $= 64$ bits, by comparing this figure with figure 5 in [7] and taking into account that the word size was divided by two (16 versus 32 respectively).

Figure 2 shows the energy consumption of RC5 that was computed from [7] and the estimation that we did for this algorithm.

RC5 and DES have a similar constant term E_0^{enc} whereas the energy per round E_r^{enc} is much higher for RC5 than for DES with a factor greater than 8. As both algorithms are based on the same basic bitwise operations, this difference can be explained by the implementation and the quality of the measures. The important point is that table II gives us a good idea of the order of energy consumption of a symmetric cryptography algorithm.

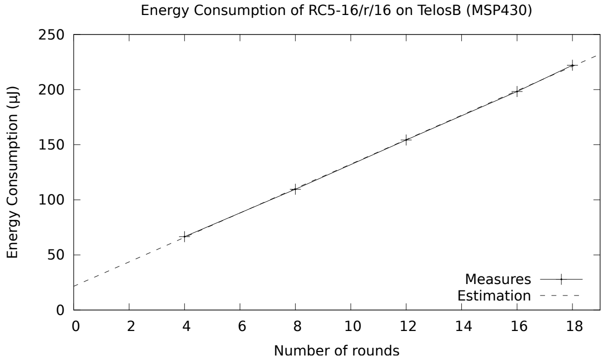


Figure 2. Energy consumption of RC5 with various number of rounds [7]

TABLE III. VALUES OF E_0^{COMP} AND COMPRESSION RATIO α FOR VARIOUS COMPRESSION ALGORITHMS

Algorithm	E_0^{comp} (nJ/bit)	α
RLE	1.325	17%
S-LZW	5.6	53%
K-RLE	2.575	56%

B. Modeling compression

Now, we model compression in the same manner as encryption. We use the results of [9] to model the energy consumption of compression as:

$$E^{\text{comp}}(\lambda) = \lambda \times E_0^{\text{comp}} \tag{2}$$

We assume that the energy consumption for compression algorithms is only proportional to the length of the input data. Generally, compression algorithms do not have a setup phase, they only take decisions according to the input data.

Table III shows the values of E_0^{comp} and the compression ratio α for various algorithms: S-LZW [11], RLE and K-RLE. S-LZW and RLE are lossless compression algorithms while K-RLE is a lossly compression algorithm. All measures are taken from [9] on the same sets of data of length $\lambda = 500$ bytes $= 4000$ bits.

C. Modeling communication

As we want to compare different scenarios with the simple scenario of just sending the data, we need a communication model. We take the communication model from [15]:

$$E^{\text{trans}}(\lambda) = \lambda \times (E_0^{\text{trans}} + \epsilon \times d^2) \tag{3}$$

In this model, E_0^{trans} is the electrical energy and is set to 50nJ/bit, ϵ is the transmit amplifier and is set to 100pJ/m^2/bit, and d is the distance to the receiving node.

V. ANALYSIS OF DIFFERENT SCENARIOS WITH COMPRESSION AND ENCRYPTION

In this section, we examine several scenarios using compression and encryption and the models described in (1), (2) and (3).

- *Scenario T*: in this scenario, we only consider the transmission of λ bits of input data.
- *Scenario CT*: in this scenario, we consider the compression of λ bits of input data with a compression ratio of α that is then transmitted.
- *Scenario ET*: in this scenario, we consider the encryption of λ bits of input data that is then transmitted.
- *Scenario CET*: in this scenario, we consider the compression of λ bits of input data with a compression ratio of α that is then encrypted and transmitted.

There is no need for a fifth scenario with encryption followed by compression and then by transmission as it would consume more energy than scenario CET.

Our goal is to show that scenario CET can consume as much energy as scenario T which would provide energy-free security.

A. Energy for the different scenarios

The energy consumption for scenario T is given by:

$$\begin{aligned} E^{\text{T}}(\lambda) &= E^{\text{trans}}(\lambda) \\ &= \lambda \times (E_0^{\text{trans}} + \epsilon \times d^2) \end{aligned} \tag{4}$$

The energy consumption for scenario CT is given by:

$$\begin{aligned} E^{\text{CT}}(\lambda) &= E^{\text{comp}}(\lambda) + E^{\text{trans}}((1-\alpha) \times \lambda) \\ &= \lambda \times (E_0^{\text{comp}} + (1-\alpha) \times (E_0^{\text{trans}} + \epsilon \times d^2)) \end{aligned} \tag{5}$$

The energy consumption for scenario ET is given by:

$$\begin{aligned} E^{\text{ET}}(\lambda, r) &= E^{\text{enc}}(\lambda, r) + E^{\text{trans}}(\lambda) \\ &= \lambda \times (E_0^{\text{enc}} + r \times E_r^{\text{enc}} + E_0^{\text{trans}} + \epsilon \times d^2) \end{aligned} \tag{6}$$

The energy consumption for scenario CET is given by:

$$\begin{aligned} E^{\text{CET}}(\lambda, r) &= E^{\text{comp}}(\lambda) + E^{\text{enc}}((1-\alpha).\lambda, r) + E^{\text{trans}}((1-\alpha).\lambda) \\ &= \lambda \times (E_0^{\text{comp}} + (1-\alpha) \times (E_0^{\text{enc}} + r \times E_r^{\text{enc}} + \\ &\quad E_0^{\text{trans}} + \epsilon \times d^2)) \end{aligned} \tag{7}$$

Table IV shows the energy consumptions with $\lambda = 64$ bits, $r = 8$ rounds and $d = 25$m in the different scenarios. $\lambda = 64$ bits is a typical size for a physical scalar data like temperature. $r = 8$ rounds is rather weak for DES and RC5. $d = 25$m is a typical distance in wireless sensor networks. We observe that, as expected, with any choice of algorithms, scenario CT consumes less energy than scenario T that consumes less energy than scenario CET that consumes less energy than scenario ET.

Table V shows the energy consumptions with $\lambda = 64$ bits, $r = 16$ rounds and $d = 75$m in the different scenarios. In this case, the number of rounds is $r = 16$, which is the maximum for DES and which is nearly the recommended number of rounds for RC5. The distance has been extended to 75m. We note that the energy consumption of scenario CET is

TABLE IV. ENERGY CONSUMPTIONS (IN μJ) WITH $\lambda = 64$ BITS, $r = 8$ ROUNDS AND $d = 25$M

Algorithms	E^{T}	E^{CT}	E^{ET}	E^{CET}
DES + RLE	7.200	6.061	37.132	30.904
DES + S-LZW	7.200	3.742	37.132	17.810
DES + K-RLE	7.200	3.333	37.132	16.503
RC5 + RLE	7.200	6.061	117.449	97.567
RC5 + S-LZW	7.200	3.742	117.449	55.559
RC5 + K-RLE	7.200	3.333	117.449	51.842

TABLE V. ENERGY CONSUMPTIONS (IN μJ) WITH $\lambda = 64$ BITS, $r = 16$ ROUNDS AND $d = 75$M

Algorithms	E^{T}	E^{CT}	E^{ET}	E^{CET}
DES + RLE	39.200	32.621	79.914	66.414
DES + S-LZW	**39.200**	18.782	79.914	**37.918**
DES + K-RLE	**39.200**	17.413	79.914	**35.327**
RC5 + RLE	39.200	32.621	238.168	197.765
RC5 + S-LZW	39.200	18.782	238.168	112.298
RC5 + K-RLE	39.200	17.413	238.168	104.959

less than the energy consumption of scenario T, with the DES algorithm combined with S-LZW or K-RLE. In the other case, the compression algorithm does not have a good enough compression ratio (RLE), or the encryption algorithm consumes so much that it cannot be counterbalanced by compression (RC5). These figures, with realistic parameters, show that it is possible to have energy-free security but we need a more precise condition.

B. Energy-free security

In this section, we try to precise the previous result, i.e., we try to compute the maximum number of rounds $r_{\max}$ for various values of d and the given compression algorithms. The following theorem gives the computation of $r_{\max}$:

Theorem 1: Security is free if and only if:

$$r \le \underbrace{\frac{\alpha \times (E_0^{\text{trans}} + \epsilon \times d^2) - (1-\alpha) \times E_0^{\text{enc}} - E_0^{\text{comp}}}{(1-\alpha) \times E_r^{\text{enc}}}}_{=r_{\max}(\alpha,d)}$$

The proof is straightforward, it directly comes from (4) and (7) with the condition that $E^{\text{CET}}(\lambda, r) \le E^{\text{T}}(\lambda)$. We note that the condition does not depend anymore on the length of the data which is normal because, in each scenario, the energy is proportional to the length of the data.

Table VI shows $r_{\max}(\alpha, d)$ for the three compression algorithms and the DES algorithm, with d varying from 25m to 100m. We observe that for distance of 25m, security can not be free whatever the compression algorithm is, i.e., $r_{\max}(\alpha, d) < 0$. The RLE algorithm do not compress enough and can never provide free security.

The results for S-LZW and K-RLE are quite close. For a distance of 50m, the maximum number of rounds is 1 and 3 respectively, which provides no security at all as there exists some easy known attacks on DES. For a distance of 75m, as already seen, the full DES with 16 rounds can be used for free with both compression algorithms. For a distance of 100m, Triple-DES, that has 48 rounds and provides strong security, can be used for free in the case of K-RLE.

Table VII shows $r_{\max}(\alpha, d)$ for the three compression algorithms and the RC5 algorithm, with d varying from 25m to 100m. RC5 consumes more energy than DES and the results for RC5 are not very good. Even for a distance of 100m, the maximum number of rounds is 4 and 5 for S-LZW and K-RLE respectively, which does not provide any security. The solution in this case is to optimize the implementation of RC5 or to find a better compression algorithm.

TABLE VI. $r_{\max}(\alpha, d)$ WITH THE DES ALGORITHM

Algorithms	25m	50m	75m	100m
RLE	–	–	–	–
S-LZW	–	1.291	18.024	41.450
K-RLE	–	3.645	22.531	48.970

TABLE VII. $r_{\max}(\alpha, d)$ WITH THE RC5 ALGORITHM

Algorithms	25m	50m	75m	100m
RLE	–	–	–	–
S-LZW	–	–	1.976	4.823
K-RLE	–	0.228	2.524	5.737

VI. ANALYSIS WITH A LINEAR NETWORK

In this section, we try to improve the previous results considering a linear network. Our idea is that the energy consumed on the sending node can be counterbalanced globally over the network by the savings of the other nodes, due to the size of the compressed data. This could improve the security of the data while still competing with scenario T.

A. Model

We use a linear network of $n + 1$ nodes, the first node that generates data and n relays in the multi-hop communication to the base station. Each node only communicates with its closest neighbors that are at distance d. The last node communicates with the base station that is at distance d too.

On this linear network, we examine the global energy in scenario T and scenario CET. In each scenario, the data is sent by the first node, then received n times and transmitted n times until the base station.

We still use the communication model from [15] for receiving:

$$E^{\text{recv}}(\lambda) = \lambda \times E_0^{\text{recv}} \tag{8}$$

E_0^{recv} is the electrical energy and is set to 50nJ/bit.

The energy consumption for scenario T with a linear network is given by:

$$\begin{aligned} E_{\text{net}}^{\text{T}}(\lambda, n) &= E^{\text{trans}}(\lambda) + n \times (E^{\text{recv}}(\lambda) + E^{\text{trans}}(\lambda)) \\ &= \lambda \times (n \times E_0^{\text{recv}} + (n+1) \times (E_0^{\text{trans}} + \epsilon \times d^2)) \end{aligned} \tag{9}$$

The energy consumption for scenario CET with a linear network is given by:

$$\begin{aligned} E_{\text{net}}^{\text{CET}}(\lambda, r, n) &= E^{\text{CET}}(\lambda, r) + n \times (E^{\text{recv}}((1-\alpha).\lambda) + \\ &\quad E^{\text{trans}}((1-\alpha).\lambda)) \\ &= \lambda \times (E_0^{\text{comp}} + (1-\alpha) \times (E_0^{\text{enc}} + r \times E_r^{\text{enc}} + \\ &\quad n \times E_0^{\text{recv}} + (n+1) \times (E_0^{\text{trans}} + \epsilon \times d^2))) \end{aligned} \tag{10}$$

TABLE VIII. $r^{\mathrm{NET}}_{\max}(\alpha, 25, n)$ WITH THE DES ALGORITHM AND A LINEAR NETWORK

Algorithms	$n = 1$	$n = 2$	$n = 5$	$n = 10$	$n = 15$
RLE	–	–	–	2.615	10.517
S-LZW	–	8.653	34.756	78.262	121.767
K-RLE	2.134	11.955	41.416	90.518	139.620

TABLE IX. $r^{\mathrm{NET}}_{\max}(\alpha, 25, n)$ WITH THE RC5 ALGORITHM AND A LINEAR NETWORK

Algorithms	$n = 1$	$n = 2$	$n = 5$	$n = 10$	$n = 15$
RLE	–	–	–	0.103	1.064
S-LZW	–	0.837	4.010	9.297	14.585
K-RLE	0.045	1.238	4.819	10.787	16.754

Now we can state an extension of theorem 1 for a linear network and compute $r^{\mathrm{net}}_{\max}(\alpha, d, n)$:

Theorem 2: Security is free in a linear network of $n + 1$ nodes if and only if:

$$r \leq \underbrace{\frac{n.\alpha.E_0^{\mathrm{recv}} + (n+1).\alpha.(E_0^{\mathrm{trans}} + \epsilon.d^2) - (1-\alpha).E_0^{\mathrm{enc}} - E_0^{\mathrm{comp}}}{(1-\alpha) \times E_r^{\mathrm{enc}}}}_{=r^{\mathrm{net}}_{\max}(\alpha,d,n)}$$

B. Results

Table VIII shows $r^{\mathrm{net}}_{\max}(\alpha, d, n)$ for the three compression algorithms and the DES algorithm, with $d = 25$m and n varying from 1 to 15. We observe that with only one-hop before the base station, security is free with the K-RLE compression algorithm, even if the number of rounds provides very weak security. For $n \geq 5$, the maximum number of rounds for S-LZW and K-RLE exceeds the number of round for DES, and for $n \geq 10$, it exceeds the number of rounds for Triple-DES. This shows that very strong security can be achieved for free on a wide network.

Table IX shows $r^{\mathrm{net}}_{\max}(\alpha, d, n)$ for the three compression algorithms and the RC5 algorithm, with $d = 25$m and n varying from 1 to 15. In this case, the situation is better than in the experiment with a single node, but the level of security is not as strong as the level for DES. For $n \geq 15$, the maximum number of rounds is 14 and 16 for S-LZW and K-RLE respectively, which a little less than the recommended 18-20 rounds for good security. Once again, the implementation of the algorithm must be improved in order to achieve better results.

VII. ANALYSIS WITH RANDOM NETWORKS

In this section, we compute $r_{\max}$ on random networks. We focus on DES and K-RLE as it's the best combination of a compression algorithm and an encryption algorithm that we have.

A. Experiment

The difficulty in this experiment is to choose an application to make the measures. We decided to test a simple routing application on a square area of width w. The network is composed of n nodes uniformly distributed on the area. Two nodes can communicate if their distance is less than $0.4 \times w$ so that there is, on average, half of the nodes in the neighborhood of each node, whatever the width of the area.

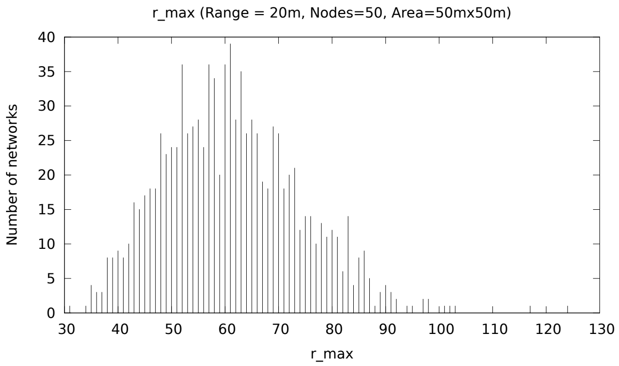


Figure 3. Distribution of $r_{\max}$ for n =50 nodes and $w = 50$m

Among the n nodes, 20 nodes are chosen to be sources of messages of size $\lambda = 64$bits. Those messages are routed to a sink which is placed at the coordinates $(0.9 \times w, 0.9 \times w)$ thanks to a shortest path algorithm which takes into account the square of the distance between each node (as the energy for transmitting a message is proportional to the square of the distance). Then, each source sends a message to the sink along the chosen path.

To compute $r_{\max}$ for a given network, we first compute the energy E_T that is consumed for scenario T. Then, we compute the energy that is consumed for scenario CET with $r = 0$ rounds, which is necessarily less than the E_T (scenario CET with $r = 0$ rounds is similar to scenario CT). Then, the number of rounds is increased until the energy is more that E_T which means we have reached $r_{\max}$.

This experiment is repeated on 1000 different random network for each value of (n, w).

B. Results

Figure 3 shows the distribution of $r_{\max}$ for n =50 nodes and $w = 50$m. We observe that this distribution is not symmetric and is quite wide so that the computation of an average is not very relevant. That's why we decided to compute the first decile of the measures, i.e., the value of $r_{\max}$ that divide the data set in 10% of low values and 90% of high values. In this case, the first decile is 45 which means that for a random network with our simple routing application, taking DES with 45 rounds (nearly Triple DES) and K-RLE is energy-free with a probability of 0.9.

This result is not a surprise as it can be compared to the results of table VIII. The range of the network is a little shorter in the case of random networks (20m), but the average number of hops from the sources to the sink must be high enough so that the energy for encryption is counterbalanced by the savings along the paths.

Table X shows the computation of the first decile for many values of (n, w). This table shows that in any case, it is possible to have strong energy-free security on a random network. We observe that, for a fixed w, $r_{\max}$ increases sub-linearly w.r.t. n. Adding more nodes on the area make paths shorter, but not short enough so that the gain in energy can bring many more rounds. We also observe that for a fixed n, $r_{\max}$ increases over-linearly w.r.t. w. In this case, the distances

TABLE X. First decile of $r_{\max}$ for a network of n nodes on a square area of width w

w n	50m	100m	200m	300m	400m
50	45	52	72	102	141
100	76	84	97	119	148
150	98	108	121	138	163
200	101	125	137	152	173

between nodes is increased and the gain in energy can be used to do many more rounds.

VIII. Conclusion

We show that it is possible to provide strong and free security thanks to a careful choice of compression algorithm and cryptographic algorithm. We provide models for the energy consumption of compression algorithms and encryption algorithms. Our models are derived from experiments done by ourselves or found in the literature, with a MSP430-based node.

It would be interesting to implement these algorithms on real nodes and check that security is really free. The consumption of the transmission comes from a model that is not derived from real experiments so an extension of this work could be to verify this model with various radio chips.

References

[1] D. Vu and D. K. Vu, “Wireless sensor network architecture and its security challenges,” Master’s thesis, California State University, Sacramento, 2010.

[2] D. Martins and H. Guyennet, “Security in wireless sensor networks: a survey of attacks and countermeasures,” International Journal of Space-Based and Situated Computing, vol. 1, no. 2, 2011, pp. 151–162.

[3] H. Saxena, C. Ai, M. Valero, Y. Li, and R. Beyah, “DSF - A Distributed Security Framework for Heterogeneous Wireless Sensor Networks,” in Military Communications Conference, 2010-Milcom 2010. IEEE, 2010, pp. 1836–1843.

[4] G. Gaubatz, J. Kaps, E. Ozturk, and B. Sunar, “State of the art in ultra-low power public key cryptography for wireless sensor networks,” in Pervasive Computing and Communications Workshops, 2005. PerCom 2005 Workshops. Third IEEE International Conference on. IEEE, 2005, pp. 146–150.

[5] X. Ren and H. Yu, “Security mechanisms for wireless sensor networks,” International Journal of Computer Science and Network security (IJCSNS), vol. 6, no. 3, 2006, pp. 155–161.

[6] A. Perrig, R. Szewczyk, J. Tygar, V. Wen, and D. Culler, “SPINS: Security protocols for sensor networks,” Wireless networks, vol. 8, no. 5, 2002, pp. 521–534.

[7] J. Lee, K. Kapitanova, and S. Son, “The price of security in wireless sensor networks,” Computer Networks, vol. 54, no. 17, 2010, pp. 2967–2978.

[8] N. Kimura and S. Latifi, “A survey on data compression in wireless sensor networks,” in Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on, vol. 2. IEEE, 2005, pp. 8–13.

[9] E. Capo-Chichi, H. Guyennet, and J. Friedt, “K-RLE: A new data compression algorithm for wireless sensor network,” in Sensor Technologies and Applications, 2009. SENSORCOMM’09. Third International Conference on. IEEE, 2009, pp. 502–507.

[10] F. Marcelloni and M. Vecchio, “A simple algorithm for data compression in wireless sensor networks,” Communications Letters, IEEE, vol. 12, no. 6, 2008, pp. 411–413.

[11] C. Sadler and M. Martonosi, “Data compression algorithms for energy-constrained devices in delay tolerant networks,” in Proceedings of the 4th international conference on Embedded networked sensor systems. ACM, 2006, pp. 265–278.

[12] K. Piotrowski, P. Langendoerfer, and S. Peter, “How public key cryptography influences wireless sensor node lifetime,” in Proceedings of the fourth ACM workshop on Security of ad hoc and sensor networks. ACM, 2006, pp. 169–176.

[13] N. (NIST), “FIPS 46-3, Data Encryption Standard (DES).”

[14] C. Burin des Roziers, G. Chelius, T. Ducrocq, E. Fleury, A. Fraboulet, A. Gallais, N. Mitton, T. Noël, and J. Vandaele, “Using SensLAB as a first class scientific tool for large scale wireless sensor network experiments,” NETWORKING 2011, 2011, pp. 147–159.

[15] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “Energy-efficient communication protocol for wireless microsensor networks,” in Proceedings of the 33rd Annual Hawaii International Conference on System Sciences. IEEE, 2000, pp. 8020–8029.

PRISMA: A Publish-Subscribe and Resource-Oriented Middleware for Wireless Sensor Networks

José R. Silva, Flávia C. Delicato, Luci Pirmez, Paulo F. Pires, Jesus M. T. Portocarrero
PPGI-DCC/IM
Federal University of Rio de Janeiro, UFRJ
Rio de Janeiro, Brazil
{jr.joserenato.jr, fdelicato , luci.pirmez, paulo.f.pires, jesus140}@gmail.com

Taniro C. Rodrigues, Thais V. Batista
DIMAp
Federal University of Rio Grande do Norte, UFRN
Natal, Brazil
{tanirocr, thaisbatista}@gmail.com

***Abstract*—PRISMA is a resource-oriented publish/subscribe middleware for WSN, which the main goals are to provide: (i) programming abstraction through the use of REpresentational State Transfer (REST) interfaces, (ii) services, encompassing asynchronous communication, resource discovery and topology control, (iii) runtime support through the creation, configuration, and execution of new applications in WSN, and (iv) QoS mechanisms to meet applications constraints. This paper describes PRISMA architecture, its implementation in the Arduino platform, and a preliminary evaluation.**

Keywords - middleware; publish/subscribe; topology control.

I. INTRODUCTION

Wireless Sensor Network (WSN) technology has been evolving fast in recent years. There are currently several hardware platforms available for WSN, such as the sensor motes manufactured by MEMSIC (former Crossbow [1]), Sun Spots [2] (now Oracle) and, more recently, Arduino platform [3], that is used in this work. Additionally, WSNs have gained a lot of attention in the research community and are becoming increasingly popular in the industry, due to their wide range of potential applications.

Early applications developed for WSN presented simple requirements and did not demand the use of complex software infrastructures. Moreover, WSN were typically designed to meet the requirements of a single target application. In other words, the source code installed in the nodes was commonly monolithic, highly tied to the requirements of a single application and to a specific sensor platform and the protocol stack for such platform. Furthermore, the application development was highly coupled to low-level primitives provided by the WSN operational system and the design approach was focused on improving the network energy efficiency, given the limited resources of nodes. Such dependence between the application layer and the underlying layers (protocols and hardware) is not desirable for emergent applications and new trends in the field, where the same physical infrastructure of a potentially heterogeneous WSN may be used for various applications, whose requirements are not known at the network deployment time [4]. As the number of WSN physical infrastructure currently deployed is increasing, there is a trend to share and integrate the sensing data produced by these networks through different applications, as well as growing initiatives to include monitoring data as part of Web applications, integrated to other types of resources available on the Internet. In such scenario, there must be interoperability between different WSNs, possibly between different applications, and between WSNs and external networks, as the Internet.

In order to meet the emerging trends of WSN scenarios, there is a need to adopt software platforms at the middleware level. A middleware can provide abstractions to build applications and to access data produced by the network, and offer generic or domain specific services. It can also provide a uniform API and standardized protocols that allow interoperability in an environment with high degree of heterogeneity. Despite of the fact that middleware platforms are widely used in traditional distributed systems, their development in the WSN context is relatively recent [4].

A WSN middleware is a layered software that lies between application code and the communication infrastructure providing, via component interfaces, a set of services that may be configured to facilitate the application development and execution in an efficient way for a distributed environment [5]. Thus, the main goal of a middleware is to enable the interaction and the communication between distributed components, hiding from application developers the complexity of the underlying hardware and network platforms, and freeing them from explicit manipulation of protocols and infrastructure services. Besides these generic requirements, a WSN middleware needs to consider some basic features, specific to this context. According to Wang et al. [4], a WSN middleware should offer four main features: (i) programming abstractions, (ii) services, (iii) runtime support, and (iv) mechanisms for Quality of Service (QoS) provision. Programming abstractions define the interface of the middleware for the application developer. Services provide implementations to achieve the abstractions; thus, services encompass the functionalities provided by the middleware and comprise the middleware core. Runtime support acts like an extension of the embedded operating system to support the middleware services. Finally, QoS mechanisms are used to meet quality constraints imposed by applications such as network lifetime, coverage, accuracy, latency, bandwidth, and others. There are several works [6]–[9] proposing middleware platforms for WSN addressing issues such as interoperability between heterogeneous devices, support for multiple application domains, adaptation and context awareness, service discovery, management of devices and other features. However, few of these works address all four requirements of WSN middleware aforementioned [4].

In this context, this paper introduces PRISMA, a resource-oriented publish/subscribe middleware for WSN, which aims to provide the aforementioned main functionalities required for WSN middleware. PRISMA programming abstraction for client applications is based on REpresentational State Transfer (REST) [10]. REST defines a lightweight communication between applications based on Web standards to facilitate the access to sensor generated data. Using a REST-based approach, the WSN, its nodes, and the sensing units of each node are described and accessed by end users and client applications as *resources*, in the same way as traditional Web resources are accessed through the Internet. By providing a high level and standardized interface for data access, PRISMA allows interoperability of networks from different technologies, thus aligning to the current trend of building heterogeneous systems involving multiple networks and applications.

PRISMA functionalities (see Section II) include (i) mechanisms to facilitate the creation and execution of WSN applications; (ii) a topology control service aiming at efficiently managing the energy consumption of nodes; (iii) capability of configuring applications QoS parameters, such as network lifetime and maximum delay; (iv) asynchronous communication via the publish/subscribe paradigm. This last is a significant feature since several WSN applications are event-driven; thus, the traditional request-reply communication model is not proper for most scenarios.

Although its logic architecture is agnostic regarding the underlying sensor platform, PRISMA physical design and implementation were tailored to Arduino-based platforms. The main motivation for using Arduino is the fact it is open-hardware, still poorly explored by the academic community of WSN, mainly in the middleware field. Moreover, this platform provides a high-level language that can be leveraged to facilitate the application development.

The rest of this paper is structured as follows: Section II presents PRISMA specification and logic architecture; Section III describes the implementation for Arduino platform; Section IV discusses related work; Section V presents performed evaluations, and finally, Section VI contains conclusions and future work.

II. PRISMA

This section presents an overview of PRISMA, its logic and physical architecture, the system operation, and the available services.

A. OVERVIEW

PRISMA assumes a heterogeneous and hierarchical WSN, with three levels: (i) Gateway, (ii) Cluster Head, and (iii) Sensor Node. The top level is represented by Gateways that are responsible for managing the network from a high level viewpoint, taking the global decisions on the system operation. In the intermediate level, Cluster Heads locally manage their respective clusters in the network. Each cluster encompasses a set of sensor nodes and one cluster leader (the Cluster Head). Cluster Heads require higher computational power than ordinary nodes since they are in charge of managing functions for a part of the network. This additional computational power is required to store information about the nodes in each cluster. Finally, in the lower layer a huge number of sensor nodes (also called as ordinary nodes) are responsible for collecting environmental data and taking local decisions. This hierarchical approach was adopted to promote scalability and facilitate network management.

PRISMA adopts REST design pattern to facilitate the access to WSN data and to support interoperability with other networks. PRISMA communication service encompasses the communication with the WSN nodes and between the middleware and external networks (as the Internet). The communication is provided by a Web Server and a broker to provide asynchronous communication. In the development of the middleware communication service, we adopted REST to communicate with client applications. In REST-based Web services, the uniform interface for accessing resources is given by Hypertext Transfer Protocol (HTTP) methods. The middleware designer has the responsibility of defining the granularity of the provided REST resource: (i) the entire WSN can be seen as a unique resource; (ii) each individual node can be exposed as a resource; and (iii) there may be many resources in each node, for example, each sensing capability (temperature, light, etc.) deployed in one single sensor node.

Besides using REST interfaces to interact with client applications, PRISMA adopts the publish-subscribe paradigm to notify its clients about events of interest. To receive notification messages, a client application must be subscribed in a publish-subscribe topic. A publish-subscribe topic is an asynchronous communication channel used by the middleware to publish interest messages to client application. This topic will be created if the application requirements can be satisfied with the WSN resources. If the WSN can meet the requirements, the client will receive the topic in response to the REST request; otherwise, will receive an error message. With this response information the client will subscribe to the desired topic in PRISMA broker and receive the data of interest whenever it is available.

B. ARCHITECTURE

PRISMA design follows a layered architecture, shown in Fig. 1, composed of three layers (i) Access, (ii) Service, and (iii) Application, described as follows. A brief description of the services provided by PRISMA software components will be presented after the description of the three layers.

- Access layer: consists of four components: **Communication, Data Acquisition, Context Monitor and Topology control.** The **Communication component** is responsible for receiving and extracting data from messages transmitted by the WSN nodes. This component includes drivers for translation and composition of messages that travel in the WSN and a listener to capture messages. The **Data Acquisition component** manages the data collection via sensing units of the nodes. The **Context Monitor component** is in charge of monitoring the network execution context in order to verify if QoS requirements are being fulfilled. An example of monitored context is the energy level of devices, which directly relates to the QoS requirement of network lifetime. The **Topology control component** is responsible for the network logical organization. This component performs the initial network configuration and a reconfiguration whenever (ii) a new application arrives, (ii) the network energy level is lower than a critical parameter, or (iii) a device failure occurs.
- Service layer: consists of three components: **Event,** responsible for managing and notifying requested events from applications in execution; **Publish and Discovery,** responsible for registering and publishing new services to be offered by the network (providing PRISMA resource discovery service); and **Decision,** responsible for analyzing arriving applications in order to verify available devices that satisfy the specified applications requirements. **Decision** is the decision-making center of the middleware (further described later).
- Application layer: consists of two components: **Application Control, Publish and Subscribe Proxy** and the **Web Server.** The first is responsible for receiving and managing applications sent to the WSN through a REST interface in a configuration file. Upon a parse of the file, its content is forward to the **Decision** component. The **Proxy Publish and Subscribe** allows asynchronous communication with client applications through the publish-subscribe paradigm. This component acts as a broker that manages the queues of PRISMA publish-subscribe implementation [11]. The **Web Server** is responsible for providing the REST interfaces that PRISMA offers, such as: (i) *Create* interface that receives new applications to be executed on the WSN; (ii) *GetServices* interface responsible for advertising the services available in the WSN; (iii) *GetData* interface, responsible for querying the data collected by the WSN (historical or current data).

These software components are divided into three subsystems, each one corresponding to a different level of the physical components considered in our architecture: (i) **Gateway** (ii) **Cluster Head** and (iii) **Sensor Nodes**. At the Gateway subsystem all components of the architecture are deployed, except the Data Acquisition component. At the Cluster Head subsystem all components of the Service Layer and Access Layer are deployed, except the Data Acquisition component that is specific to the Sensor Node subsystem. At the Sensor Node subsystem all components of the Service Layer and Access Layer are deployed. The only subsystem that communicates with client applications is the gateway subsystem for being the one that includes all the components of the Application Layer.

PRISMA provides four services: (i) communication; (ii) topology control; (iii) resource discovery; and (iv) context monitoring. The communication service is responsible for the communication among middleware components and the WSN nodes, and with external entities (client applications or the Internet). This service is provided by the following

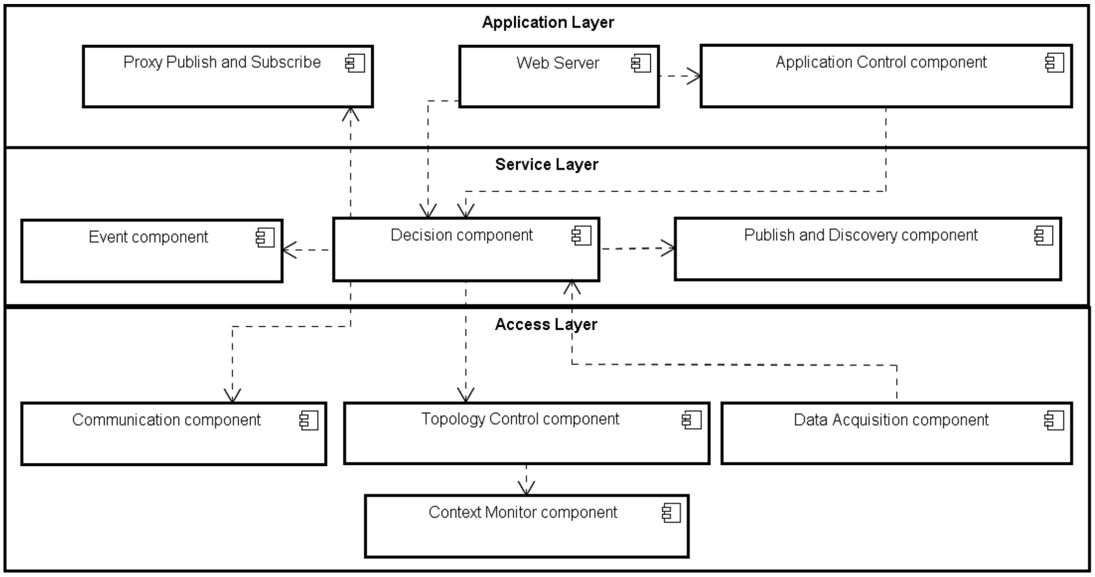

Application Layer
Proxy Publish and Subscribe
Web Server
Application Control component
Service Layer
Event component
Decision component
Publish and Discovery component
Access Layer
Communication component
Topology Control component
Data Acquisition component
Context Monitor component

Figure 1. UML component diagram illustrating the PRISMA architecture

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-360-5

components: communication component to exchange messages with the WSN; Web Server to communicate with external networks and Proxy Publish and Subscribe to communicate asynchronously with client applications. The topology control service is responsible for selecting the clusters and nodes that will participate in the sensing tasks for a given application; this service is provided by the topology control component. The resource discovery service is provided by the Publish and Discovery component. This service is responsible for identifying new nodes in the network and publishing new available services to the decision maker center of the middleware. The context monitor service is provided by the Context Monitor component and it is responsible for monitoring the energy of WSN/Cluster/Node (depending on the subsystem) and detecting conditions of lacking of energy.

C. System Operation

This subsection presents PRISMA operation from the sensor nodes deployment to the creation and configuration of applications on WSN nodes. The UML diagram activity of Figure 2 depicts the main steps of this operation and following we briefly describe these steps.

Initially, the middleware (software) components are installed on physical devices according to their functionality (**Cluster Heads** or **Sensor Nodes**). Information about the geographical area of deployement and the Cluster Head assigned for each node are “hard-coded” into the code of the nodes. Then, considering that the nodes are distributed in their respective target areas, the **Resource Discovery** service starts. This service is responsible for identifying each sensor node that is active in a specific geographic area as well as its sensing capabilities (the provided services/resources). The process starts with the sensor node sending a message to their respective Cluster Head. Then, the **Cluster Head** updates its node list and sends a message to the Gateway containing the description of the set of sensing capabilities managed by the (new) nodes to the decision-making center of the middleware. This message includes, for each sensor node, the following information:

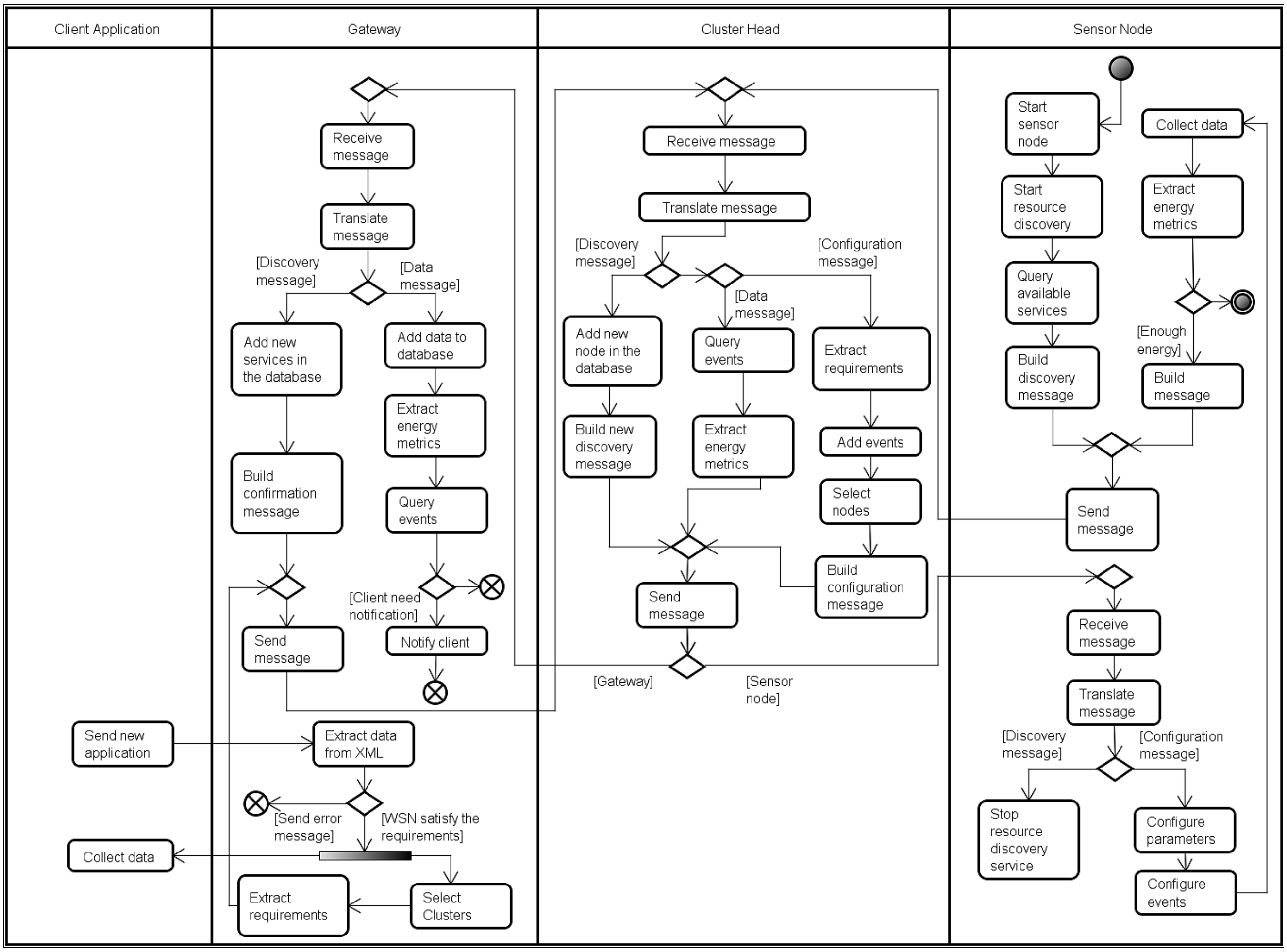
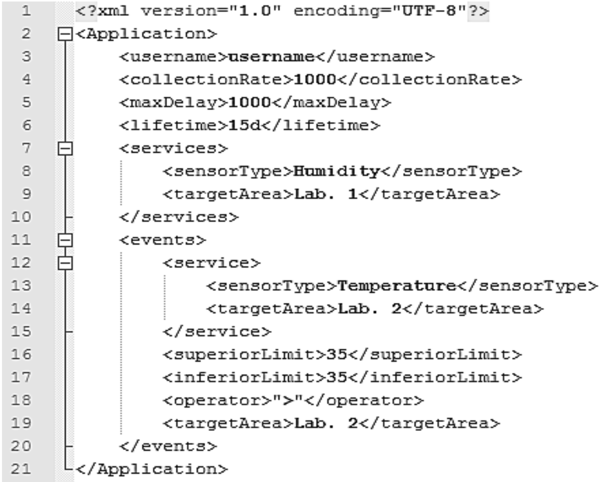


Figure 2. UML Activity diagram of PRIMA operation

(i) address defined in the sensor node radio (MAC address), (ii) cluster to which it belongs (Cluster ID), (iii) available resources (sensing units, as for example, temperature, humidity) and (iv) residual energy. The **Gateway** then updates its database with this information using the **Publication and Discovery** component. The **Context Monitor** is responsible to identify when a node's energy is almost depleted and notify the Resource Discovery service to advertise the Cluster Head and Gateway of its low energy in the same process described above.

Once the network is organized, all its resources are listed and made available through the **Gateway**, applications can be created in the WSN. The Gateway receives new applications to be deployed in WSN through a REST interface. Hence, in PRISMA all access to the WSN resources and creation of new applications follows a RESTful approach. Configuration files are submitted via a REST interface to create applications and each file is translated in parameters to configure the network. These parameters are sent to specific clusters, which are responsible for organizing themselves in order to provide required data and services. This organization comprises the selection of the nodes to actively participate in data collection following the requirements sent by the application. Applications can be created through the *Create* REST interface (one of the REST interfaces available on PRISMA by the **Web Server** component). Such interface receives an eXtensible Markup Language (XML) [12] file through a HTTP POST message [13], in the following url: http://ServerAddress:8080/prisma/rest/applications/create. Figure 3 depicts an example of a configuration file.

The process to create a new application starts by receiving a XML configuration file (Figure 3) that specifies the application requirements to execute into the nodes (sent by the client application); this XML file is translated in one Java object by the **Application Control** component and, after that, these requirements are verified to define whether the required services (specific ability of every sensor node) may be attended by one or a set of available sensor nodes, the **Decision** component query the Publish and Discovery component to check the available services to verify this. After that, requirements are translated to parameters to be configured in the selected nodes to meet the specified requirements. A configuration file may have many services where each one defines a sensing task. A XML configuration file defines: (i) a periodic application, for instance, to monitor temperature of an environment every 5 minutes; (ii) an event-driven application, for instance, to monitor temperature in an environment and to notify the client whenever a sensor detects a value of 50ºC or more; or (iii) both types of applications, for instance, to monitor temperature of an environment every 5 minutes and to notify when a temperature achieve 50ºC in order to detect fire. PRISMA middleware supports multiple applications by reusing WSN data or allocating new nodes to be active.



```
1   <?xml version="1.0" encoding="UTF-8"?>
2   <Application>
3       <username>username</username>
4       <collectionRate>1000</collectionRate>
5       <maxDelay>1000</maxDelay>
6       <lifetime>15d</lifetime>
7       <services>
8           <sensorType>Humidity</sensorType>
9           <targetArea>Lab. 1</targetArea>
10      </services>
11      <events>
12          <service>
13              <sensorType>Temperature</sensorType>
14              <targetArea>Lab. 2</targetArea>
15          </service>
16          <superiorLimit>35</superiorLimit>
17          <inferiorLimit>35</inferiorLimit>
18          <operator>">"</operator>
19          <targetArea>Lab. 2</targetArea>
20      </events>
21  </Application>
```

Figure 3. New application configuration file

Figure 3 describes an application ready to be executed in the network. It is possible to define the following requirements for an application: (i) data collection rate (in milliseconds), (ii) maximum delay (in milliseconds), and (iii) the application lifetime, which can be set in hours or days. In addition to these requirements, this configuration file also defines which services are required and in which geographical area these services should be located. The target areas are statically configured (at pre-deployment time) by the WSN administrator in the sensor nodes. Between lines 7 and 10 of the example we define the service to monitor **humidity** data in the area. This data should be collected from the **Lab. 1** (a symbolic region) and respect the collection rate set in line 4 (1 second).

We can also define events that will be monitored by setting the service, target geographic area, upper/lower limit when applicable and the comparison operator being used. The **Event** component is responsible for recording these events and checks them each time new data is received. Lines 11 to 20 specify to collect **temperature** data and notify (send data messages) only when they are higher than 35 degrees in the geographical area named **Lab 2**. The publish-subscribe topic for this application is created by the **Proxy Publish and Subscribe** after receiving this configuration file. The access information to the topic (the topic name) is sent as a response through the REST interface accessed by the client. The client subscribes this topic to receive the asynchronously requested data.

The **Decision** component is responsible for analyzing the arriving sensing applications, extract its requirements and query the **Topology Control** component to verify the existence in the WSN of devices (nodes) that meet the requirements specified in the received configuration file. If any device meets the requirements specified, the **Decision** component will create a message and the **Communication** component will send the application requirements to the Cluster Head(s) of the respective devices. Only clusters that are selected for the execution of the application will receive these requirements. Each cluster is only assigned with tasks that can be met by its currents resources, thus avoiding sending unnecessary messages. In PRISMA, a task corresponds to the act of executing a service (for instance, a temperature sensing task) and a service corresponds to a given capability of a node (for instance, the capability of sensing temperature values).

When the requirements are received in the **Cluster Head** the local process of selecting active nodes is started (more details in Section D). After the **Topology Control** component selects the nodes that will participate in the data acquisition for the application, the received task requirements are then translated into parameters by the **Decision** component to be configured and sent to the selected nodes. These parameters are directly extracted from the configuration file: sensing capabilities required, collection rate, maximum delay, application lifetime and threshold for sending data (representing the detection of an event of interest). The selection of active nodes takes into account the residual energy of nodes, information that is updated whenever a message is sent by the node.

D. Topology Control

A major goal in the management of nodes in WSNs is to rationalize the use of energy resources of the network in order to prolong its lifetime and consume energy evenly across the nodes. One way to achieve this goal is by adopting a scheme of rotation of the work performed by the network nodes, changing their operating mode (active or sleep mode), where the subset selected to remain active should be able to meet the requirements requested by the application.

The problem of selecting active nodes can be expressed as the algorithm that decides which sensors must remain active for a given application task. In PRISMA, this is the responsibility of the topology control algorithm, the core of the middleware topology control component. The algorithm proposed in this paper is based on [14], where the time is divided into *j* rounds during which the selected subset of nodes remains constant. A task starts running at the beginning of a round and can last for a time equal to an integer multiple of *p*, where *p* is the length of a round.

The mechanism of topology control is first executed when the requirements of an application are sent to the network. These requirements contain a description of the application that will run on the WSN and the desired QoS requirements. After the execution of the topology control mechanism the first round for the application in question starts. The selection algorithm can be run again in the following cases: (i) on demand by the application to change any parameters of QoS if needed, (ii) in a proactive way by the network, for the purpose of energy conservation, for instance, and (iii) reactively by the network, when the **Context Monitor** component detects a QoS requirement is not being met.

Unlike the proposal of the algorithm described by Delicato et al. [14], that is based on a flat network of homogeneous sensors, and where the process of selecting active nodes is centralized, in PRISMA the network is heterogeneous, and a hierarchical selection is performed in two levels: the first level corresponds to the global view of the network and second level corresponds to the local view of the network, within each cluster. PRISMA approach has the potential of allowing greater scalability since it is performed at two levels, thus being more suitable for scenarios of large-scale and shared WSNs. The algorithm does not need to flood the network to determine active nodes for every application to be executed.

By adopting a hierarchical approach, in PRISMA the topology control mechanism runs on two physical components: (i) **Gateway,** and (ii) **Cluster Head**, where each component performs the algorithm on a different level. The first level corresponds to the global view of the network (**Topology Control Component** of the **Gateway**) where the **Gateway** is responsible for a pre-selection of clusters that contain sensor nodes potentially useful for an application. At this level, the **Gateway** runs the steps of the topology control algorithm to determining the clusters to be used for a given application: clusters that do not have resources or capabilities to suit a given application will be excluded from active nodes selection process. The exclusion of these clusters is based on a simple set of rules, for example, exclusion of clusters not having the necessary services or those outside the desired geographical area. The second level, local within each cluster, will run in the **Cluster Heads** that have a higher processing power than the ordinary sensor nodes. The higher processing power and memory capacity is desired to maintain in the Cluster Head the list of nodes that are in its coverage area and information about these nodes, such as available services, energy level and its state (sleep or active). The process of changing the operating mode of the node is implemented through configuration/control messages responsible to set the duty cycle of each node.

After the pre-selection executed by the Gateway, the algorithm for selecting active nodes is triggered in a Cluster Head (selected in the first level of the topology control mechanism) whenever it receives a request to create a new application, or when the energy level of any node is below a minimum threshold for the execution of their tasks. Such algorithm running at the cluster heads receives as input the application requirements (data from the XML file) and the set of available services on its cluster, and produces as output the set of nodes to be used by the application.

The process of selecting active nodes begins with a query to retrieve the energy levels of the cluster nodes, maintained by the **Cluster Head** responsible for a given area. After running the algorithm for the selection of active nodes a message is sent to each node in the cluster to determine whether the node is active or in sleep mode. If it is in sleep mode, the next time the node wake up it will query its cluster head to check pending messages and then receive a control message. Information about energy stats is collected and sent along with the data collected by the sensors in order to supply its Cluster Head with the information on current energy levels of the network.

III. IMPLEMENTATION

The components of the **Gateway** subsystem were developed on J2EE 1.4 platform and implemented using: Apache TomEE [15] as application server, Jersey for the creation of REST interfaces; Log4J for handling logs (for debugging purposes), Hibernate to implement the persistency layer, and MySQL relational database management system [16] as the data repository. The project was developed following the design pattern Data Access Object (DAO) [17]. Asynchronous communication was developed using ActiveMQ [18] which is included in Apache TomEE. The source code can be found in the URL: http://ubicomp.nce.ufrj.br/ubicomp/projetos/prisma/.

As stated in Section I, the target sensor platform for PRISMA implementation is Arduino. The main motivation for this choice is that this is a recent platform (launched in 2005), open hardware, not explored by the academic community of sensor networks, especially in the area of middleware. To the best of our knowledge, there is currently no WSN middleware implementation in this platform reported in the literature up to date. Furthermore, the platform has a high level language which facilitates development of applications.

TABLE 1. COMPARISON TABLE OF ARDUINO MODELS

	Arduino UNO	Arduino MEGA
Microcontroller	ATmega328	ATmega1280
Operating Voltage	5V	5V
Recommended input voltage	7-12V	7-12V
Input voltage limit	6-20V	6-20V
Digital input and output pins	14 (6 can provide power)	54 (15 can provide power)
Analog input pins	6	16
Current output for I / O pins	40mA	40mA
Pin 3.3V current output	50mA	50mA
Flash memory	32KB (0.5KB is used by the bootloader)	128KB (4KB are used by the bootloader)
SRAM	2KB	8KB
EEPROM	1KB	4KB
Clock speed	16MHz	16MHz

The components of the **Cluster head** and **Sensor node** subsystems were developed using the Arduino IDE development that is available at Arduino official web site. The nodes were programmed in Arduino Programming Language [19] (based on Wiring programming language [20]). This language has three main categories of code constructs: structures, values (variables and constants) and functions. Such a language is based on C / C++ [21]. Given this fact, any function of these languages can be used in Arduino programming. As previously mentioned, PRISMA works with a heterogeneous network, where the Cluster Heads need more computing power. Therefore, Arduino MEGA was chosen for this function since it has higher computational power. The sensor nodes use the Arduino UNO model. Additional hardware details of these models can be found in Table 1. The platform offers the concept of shields, which are cards that can be added to the Arduino board to increase its functionality. There are Arduino shields for connecting with Bluetooth, Ethernet modules, among others. The shield used in this work, called XBee Shield allows the interconnection of the Arduino XBee radio module [22].

PRISMA has a set of libraries representing the following features of the middleware: topology control, services discovery, and a library for message handling. In addition to the libraries specifically developed to implement PRISMA, the following existing libraries are used: XBee-Arduino [23], responsible for communicating with the XBEE radio and PString [24] to facilitate the use of the API functions and so reduce the complexity of handling messages.

XBee radio works with two operating modes for data transmission and reception. In the first mode, called Transparent Operation or **AT**, the data is sent and received directly through the node serial port. In order to send data and use AT commands, the application code installed on nodes needs to connect to the serial port of the XBee module. Through AT commands it is possible to modify the XBee configuration of the node. This mode although simple is not scalable to send data to multiple recipients and in order to change the XBee radio configuration it is necessary to access to the physical device directly.

The second mode, which is used in this work, is the Application Programming Interface (API) mode, based on sending and receiving data frames by specifying how commands, command responses and messages about the operating status of the XBee module are sent and received. This mode allows remotely sending settings (AT commands) for the XBee radio of the nodes. By using the API mode, new nodes can be inserted in the WSN and configured on the fly, thus facilitating scale up the network. AT commands can also be sent and received via the API mode, allowing the coexistence of the two modes in one network. By using the Arduino in conjunction with the XBee radio for wireless communication it is possible to encapsulate the data exchanged over the network in packets that follow the IEEE 802.15.4 standard [25].

IV. RELATED WORK

This section analyzes existing publish-subscribe middleware platforms for WSN and compares them with PRISMA. In [26], TinyDDS is described as a (re)configurable and open source middleware, developed using design patterns, and aimed at offering interoperability of publish-subscribe WSN applications. TinyDDS addresses most of the features mentioned in Section I, except the QoS mechanism. Moreover, TinyDDS does not consider adaptation issues and does not include a topology control mechanism. In this sense, PRISMA provides is a more comprehensive middleware solution.

MiSense [27] is a service-oriented and component-based middleware designed to support distributed applications running in sensors with different performance requirements. MiSense covers all the features described in Section I, but the services offered by MiSense are simple, e.g., (i) the topology control mechanism merely elects cluster heads based on their energy levels and makes these nodes responsible for forwarding all messages originated within that cluster to the sink node, overloading the cluster heads and depleting the energy quickly in a high workload condition. On the other hand, the algorithm used in PRISMA considers the application requirements to build the WSN logical topology, thus its solution tries to balance between optimization of the networks resources and the needs of final users. Also, in [27] (ii) there is an asynchronous communication service where each node has its own Broker that manages the topics and subscriptions. However, this approach based on Broker can overload nodes with high workloads because each node consumes most of its available battery transmitting the new collected data for all subscribed applications/nodes. In PRISMA, the Gateway takes those responsibilities and works as an intermediary between applications and sensor networks, avoiding excessive exchange of messages by the use of the publish-subscribe mechanism.

MufFIN [9] (*Middleware For the Internet of thiNgs*) is a IoT middleware that uses SOA principles and Sensor Web Enablement (SWE) to provide an abstraction layer to client applications. The main contributions of MufFIN are: (i) providing programming abstraction to applications and (ii) management of collected data and its broadcast to applications via Web services respecting the SWE specifications. MufFIN provides abstraction of code deployment, communications and hardware of smart objects. The authors have chosen to accommodate the differences between the heterogeneous devices at the middleware level. From the perspective of the application, all devices are reprogrammable and can communicate. MufFIN allows all its connected devices to be reprogrammable even though the device does not have this capability natively. In order to enable this feature, the middleware creates a filter (called Data-Flow) to process the information received from the WSN. For the application such approach works as if the device is running the code, but in fact the data collected pass through the Data-Flow provided by the middleware and only after such process it is delivered to the client. Differently from PRISMA, MufFIN does not provide any support for QoS management. By encompassing a topology service that also works as a QoS mechanism, PRISMA aims at providing a more complete solution, at the middleware level, for WSNs.

Mires [28] is a middleware based on both service and publish-subscribe paradigms that operates above the TinyOS layer encapsulating its interfaces and providing high-level services to applications. Internally, Mires consists of a publish-subscribe service, a routing component and additional services. Although Mires adopts a service-based design and provides asynchronous communication, it does not offer programming abstraction, runtime support or QoS mechanisms. The only mechanism to save energy included in Mires consists in reducing the number of messages sent in the network. This is accomplished by sending only messages related to subscribed topics. In PRISMA the energy is saved by the topology control algorithm. In PRISMA algorithm the requirements of client applications are checked and only target clusters and nodes within the interest area and able to provide useful service for the application will receive messages. The configuration message to subscribe to a topic will not be broadcast over the network but will be forwarded only to nodes that are active and relevant to the topic.

MARINE [29] is a component-based middleware specifically designed for WSNs, which adopts REST and microkernel architectural patterns in its design. MARINE provides a communication service based on REST and, to deal with the dynamic environment and the need for resource optimization in WSNs, it provides inspection, adaptation and configuration services. New services can be specified by third parties and incorporated using the component model and programming interfaces provided by MARINE. The asynchronous communication service is provided by the PubSubHubbub protocol. MARINE creates a Hub on each sensor node so that every request to the node is taken directly to it, creating a high energy consumption since nodes are directly accessed. MARINE provides all the functionality required by a middleware for WSN. The main difference for PRISMA is that all requests pass through the gateway that is responsible for forwarding the request to a node that can provide the data and that has enough energy to complete the task avoiding the large energy consumption mentioned above. The gateway is responsible for determining which node should answer the request.

V. EVALUATION

In all experiments performed with PRISMA, the WSN comprised of Arduino Uno sensor platform that have 2KB RAM and 32KB of flash memory for program storage. This platform is powered by four AA (1.5V, 1500mAh) batteries that provide approximately 32 kJ of energy. The PRISMA was implemented using Arduino programming language. Experiments with real sensors were performed in the Ubiquitous Computing Laboratory of PPGI-UFRJ.

Considering the objectives of this work and following the methodology goal, question, metric (GQM) [30], we defined two goals. Goal 1 (G1): Analyze PRISMA with the purpose of evaluating its effectiveness with respect to meeting the programming abstraction feature for WSN middleware in the context of application development and implementation. Goal 2 (G2): Analyze PRISMA with the purpose of evaluating its scalability in terms of the increase of application requests.

These goals were refined in four questions. Question Q1 is related to goal G1 and questions from Q2 to Q4 are related to goal G2. Q1: How expensive is it to build an application using PRISMA, in terms of lines of code? Q2: Does PRISMA scale well to serve a growing number of application requests? Q3: What is PRISMA overhead in terms of control/configuration messages? Q4: What is PRISMA overhead in terms of required RAM for its operation within WSN nodes?

The following metrics were defined to answer the questions considered in the evaluation. Each metric is denoted by M_{ij}, where *i* correspond to the question identifier, and *j* is a counter when there is more than one metric per question. The **number of lines of code (M_{11},)** is a metric used to evaluate how simple it is to create a sensing application using the abstractions provided by PRISMA (Q1). For computing this metric, we collected the number of lines of code required to create an application: (i) directly using Arduino programming and (ii) using PRISMA approach. The **Maximum number of requests supported (M_{21}):** is a metric used to assess whether PRISMA is scalable with respect to its programming abstraction approach, namely the use of REST (Q2). The **Time spent to deploy a new applications when PRISMA Web Server is overloaded (M_{22})**: is a metric used to assess the Gateway response time when a new configuration file is sent via the *Create* REST interface in a situation where many requests are made simultaneously. An increasing number of requests per second for the middleware interfaces were generated to determine the maximum number of requests supported and the delay expected to create new applications by varying the number of requirements sent to the middleware and thereby generating messages of varying size sent to the WSN. **The size of control message transmitted inside the WSN (M_{31})** is a metric is used to evaluate the overhead introduced by the control messages disseminated in the WSN (Q3). The RAM metric (**M_{41}) is used by sensing applications:** this metric is used to evaluate the overhead introduced by PRISMA (Q4). It verifies the RAM consumption when we use PRISMA to configure a sensing application.

A. *Evaluation Methodology and Scenarios*

To collect data to answer the questions an experimental evaluation was conducted. In the experiment, the network was planned so that it had two (2) clusters, each one containing a set of three (3) sensor nodes with different sensing capabilities, enabling the use of the topology control service at different times. A circular topology was organized where the sink node was at the center of the region, so that the sensor nodes and clusters remain equidistant from the center, thereby reducing the distance factor in latency and power consumption of both clusters. The radios of the sensor nodes were configured in order to respect the aforementioned topology. Nodes were hard-coded associated to their respective cluster heads and the transmission power was set as the same for all of them. Initially, all nodes had the same duty cycle. This cycle will only be changed by requests from client applications. Four client applications were developed, one producing periodic data requests and the other event-based data requests in order to collect the metrics specified. Each application represents a scenario. In the first scenario the application requests a periodic sample of temperature in the room “Lab1”. The samples are to be collected every 15 seconds and the application will remain active for 5 minutes. The second scenario specifies an application that will execute for 5 minutes and collect periodic samples of the temperature, humidity and photo sensors. These samples will be collected every 15 seconds. The third scenario specifies an application that will execute for 10 minutes and collect samples of temperature every 15 seconds. However, in this case data will be sent only if the temperature exceeds 30ºC; moreover, the application requests a maximum delay of 200ms. The fourth scenario specifies an application that will execute for 10 minutes and collects temperature and photo samples every 15 seconds. Data will be sent only if the temperature: either exceeds 40ºC or is below 15ºC. The photo data will be sent by the nodes if the luminosity of the room exceeds 600 lumens. The maximum delay for this application is defined as 300ms.

B. *Analysis of results*

This section discusses the results obtained by extracting the metrics. The results are presented in Table 2. Regarding goal 1 (**G1**), the results of the metric **M_{11}** indicate, as expected, that the programming abstraction provided by PRISMA (creation of applications via an XML file and submission through REST interfaces) makes it simple to create new applications. In addition to reducing the number of lines needed to create an application, the client uses a higher-level language to specify the requirements. It is noteworthy that the difference in the number of lines increases with the complexity of the application created. In the table, A denotes Arduino and P mean PRISMA.

TABLE 2. EVALUATION OF RESULTS USING GQM

			Periodic				Event			
Goal	**Question**	**# Services** **Metric**	**1**		**3**		**1**		**3**	
G1	**Q1**	$\mathbf{M_{11}}$ **(lines)**	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**
			15	11	37	19	20	16	50	34
G2	**Q2**	$\mathbf{M_{21}}$ **(# requests)**	1200		950		1100		890	
		$\mathbf{M_{22}}$	111 ms / 114 σ		190 ms / 172 σ		186 ms / 153 σ		486 ms / 239 σ	
	Q3	$\mathbf{M_{31}}$	74 Bytes		171 Bytes		82 Bytes		195 Bytes	
	Q4	$\mathbf{M_{41}}$ **(bytes)**	Uno	Mega	Uno	Mega	Uno	Mega	Uno	Mega
			13332	14918	13332	14918	13332	14918	13332	14918

As for goal 2 (**G2**), the result of $\mathbf{M_{21}}$ metric indicates the maximum number of simultaneous requests before PRISMA Web Server component stops responding or demonstrates an unacceptable response time. We considered any response time above 800 milliseconds as unacceptable, following literature recommendations for typical Web applications. The result of $\mathbf{M_{22}}$ metric indicates the response time for deploying new applications through the *Create* REST interface provided by PRISMA. The response time was 246 milliseconds on average. Such response time in the literature is considered an imperceptible time from the point of view of typical Web applications. The response time for the creation of event-based applications is greater than the response time for creating periodic applications due to the higher number of transactions in the database. $\mathbf{M_{31}}$ and $\mathbf{M_{41}}$ metrics assess the overhead introduced by PRISMA. $\mathbf{M_{31}}$ measures the number of bytes transmitted in WSN nodes to create an application in each of the scenarios presented. We observed that the amount of bytes sent to the WSN to configure a new application increases according to its complexity. This increase is related to the number of messages that must be exchanged for the configuration of this new application. In the worst case, one message for each event/service requested is required. This happens due to the need of sending configuration messages to the cluster head that will select nodes that participate in the application and send this new configuration for these nodes. With respect to metric $\mathbf{M_{41}}$, the table shows the RAM consumption when using PRISMA on the Arduino UNO and Arduino MEGA. We can verify that the RAM consumption did not change between the scenarios and changed only between models. The variation between the models is due to the size of the bootloader of each model. It is worth noting that PRISMA consumed less than 50% of the available RAM on the Arduino UNO (32Kb) indicating that new services and features can be added to PRISMA.

Analyzing the results, we conclude that the complexity of the application affects the size of the XML necessary to create this application and the size of messages that are transmitted on the WSNs to configure this application. In contrast, PRISMA allows creating applications on the fly. This avoids redeploying the source code on the sensor nodes each time a new application arrives. The maximum number of supported requests and delay perceived by the customer are mainly affected by the characteristics of the hardware that was used to test and to implement the gateway.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we presented PRISMA, a resource oriented, publish/subscribe middleware for Wireless Sensor Networks. Results of a preliminary evaluation demonstrated the feasibility of implementing PRISMA in real sensor nodes and shown that it provides a suitable programming abstraction for WSN application development. This result points out that our approach is a "ready to use" middleware for an easy access, recent and open-hardware WSN platform. Moreover, the use of PRISMA architecture and REST interfaces allows future developers to continue evolving this approach by creating new services or clients to access data published by a WSN using PRISMA. For future works, we intent to evaluate the remaining features of PRISMA; in particular, its QoS mechanism that is basically provided by the topology control, as well as the asynchronous communication model introduced in this paper. We also plan to perform a comparative analysis with results obtained by Mires, and to analyze the impact of various parameters on PRISMA performance (e.g., number of sensor nodes, topology and application requirements). Finally, we intend to add support for actuators to cover a wider range of possible applications to use PRISMA.

ACKNOWLEDGMENT

This work was partially supported by Brazilian Funding Agencies FAPERJ, CNPq and CENPES.

REFERENCES

[1] TinyOS, “TinyOS.” [Online]. Available: http://www.tinyos.net/. [retrieved: May, 2014].

[2] Oracle, “Oracle.” [Online]. Available: http://www.oracle.com/br/index.html. [retrieved: May, 2014].

[3] Arduino, “Arduino.” [Online]. Available: http://arduino.cc/. [retrieved: May, 2014].

[4] M.-M. Wang, J.-N. Cao, J. Li, and S. K. Dasi, “Middleware for Wireless Sensor Networks: A Survey,” J. Comput. Sci. Technol., vol. 23, no. 3, 2008, pp. 305–326.

[5] S. Hadim and N. Mohamed, “Middleware: Middleware Challenges and Approaches for Wireless Sensor Networks,” IEEE Distrib. Syst. Online, vol. 7, no. 3, Mar. 2006.

[6] X. Koutsoukos, M. Kushwaha, I. Amundson, S. Neema, and J. Sztipanovits, “OASiS: A service-oriented architecture for ambient-aware sensor networks,” Compos. Embed. Syst. Sci. Ind. Issues, vol. 4888, 2007, pp. 125–149.

[7] A. Taherkordi, Q. Le-Trung, R. Rouvoy, and F. Eliassen, “WiSeKit: A Distributed Middleware to Support Application-level Adaptation in Sensor Networks,” in Proceedings of 9th IFIP Int. Conf on Distributed Applications and Interoperable Systems (DAIS), 2009, vol. 5523, pp. 44–58.

[8] P. Boonma and J. Suzuki, “BiSNET: A biologically-inspired middleware architecture for self-managing wireless sensor networks,” Comput. Networks, vol. 51, no. 16, Nov. 2007, pp. 4599–4616.

[9] B. Valente and F. Martins, “A Middleware Framework for the Internet of Things,” Conf. Adv. Futur. Internet, no. c, 2011, pp. 139–144.

[10] R. T. Fielding, “Architectural Styles and the Design of Network-based Software Architectures,” PhD Thesis University of California, Irvine, 2000.

[11] P. T. Eugster, P. A. Felber, R. Guerraoui, and A.-M. Kermarrec, “The Many Faces of Publish/Subscribe,” ACM Comput. Surv., vol. 35, no. 2, 2003, pp. 114–131.

[12] XML, “XML.” [Online]. Available: http://www.w3.org/XML/. [retrieved: May, 2014].

[13] HTTP, “HTTP.” [Online]. Available: http://www.w3.org/Protocols/. [retrieved: May, 2014].

[14] F. Delicato, F. Protti, L. Pirmez, and J. F. de Rezende, “An efficient heuristic for selecting active nodes in wireless sensor networks,” Comput. Networks, vol. 50, no. 18, Dec. 2006, pp. 3701–3720.

[15] Apache TomEE, “Apache TomEE.” [Online]. Available: http://tomee.apache.org/apache-tomee.html. [retrieved: May, 2014].

[16] MySQL, “MySQL.” [Online]. Available: http://www.mysql.com/. [retrieved: May, 2014].

[17] Data Access Object, “Data Access Object.” [Online]. Available: http://www.oracle.com/technetwork/java/dataaccessobject-138824.html. [retrieved: May, 2014].

[18] ActiveMQ, “ActiveMQ.” [Online]. Available: http://activemq.apache.org. [retrieved: May, 2014].

[19] Arduino Programming Language, “Arduino Programming Language.” [Online]. Available: http://arduino.cc/en/Reference/HomePage. [retrieved: May, 2014].

[20] Wiring, “Wiring.” [Online]. Available: http://wiring.org.co/. [retrieved: May, 2014].

[21] D. M. Ritchie, “The development of the C language,” in The second ACM SIGPLAN Conf. on History of programming languages - HOPL-II, 1993, vol. 28, no. 3, pp. 201–208.

[22] Digi International, “XBee.” [Online]. Available: http://www.digi.com/xbee/. [retrieved: May, 2014].

[23] A. Rapp, “XBee-Arduino.” [Online]. Available: https://code.google.com/p/xbee-arduino/. [retrieved: May, 2014].

[24] M. Hart, “PString.” [Online]. Available: http://arduiniana.org/libraries/pstring/. [retrieved: May, 2014].

[25] IEEE, “802.15.4.” [Online]. Available: http://www.ieee802.org/15/pub/TG4.html. [retrieved: May, 2014].

[26] P. Boonma and J. Suzuki, “TinyDDS: An Interoperable and Configurable Publish/Subscribe Middleware for Wireless Sensor Networks,” in Wireless Technologies: Concepts, Methodologies, Tools and Applications, A. M. Hinze and A. Buchmann, Eds. IGI Global, 2011, pp. 819–846.

[27] K. K. Khedo and R. K. Subramanian, “A Service-Oriented Component-Based Middleware Architecture for Wireless Sensor Networks,” J. Comput. Sci., vol. 9, no. 3, 2009, pp. 174–182.

[28] E. Souto et al., “Mires: a publish/subscribe middleware for sensor networks,” Pers. Ubiquitous Comput., vol. 10, no. 1, Oct. 2005, pp. 37–44.

[29] F. C. Delicato et al., “MARINE : MiddlewAre for Resource and mIssion oriented sensor NEtworks,” Mob. Comput. Commun. Rev., vol. 17, no. 1, 2013, pp. 40–54.

[30] V. Basili, G. Caldiera, and H. Rombach, “The goal question metric approach,” Encyclopedia of software Engineering, vol. 2, 1994, pp. 528–532.

A New Performance Efficient Trend of Delivery Mechanism Applied to DTN Routing Protocols in VANETs

Joao Goncalves Filho,
Joaquim Celestino Jr.
Computer Networks and Security Laboratory (LARCES)
State University of Ceara (UECE)
Fortaleza, Brazil
{joao.goncalves, celestino}@larces.uece.br

Ahmed Patel
Software Technology & Management Research Center
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia (UKM)
Bangi, Sengalor, Malaysia
whinchat2010@gmail.com

***Abstract*—There are major challenges in establishing effective communications between nodes in vehicular ad hoc networks (VANETs) that are subject to disconnections, hinder end-to-end source and target connection. Another problem arises when VANETs are sparse, whereby communication between vehicles occurs after long periods of time causing delays. In these environments, traditional routing protocols proposed for VANETs suffer holding continuous connection and performance problems. To overcome these problems, Delay Tolerant Networks (DTN) for Interplanetary Networks (IPN) routing protocols, which encourages applications to use a minimized number of round trips are considered suitable alternatives. They are designed for storing and forwarding messages when nodes can find other nodes to maintain end-to-end connections. In our previous work, we proposed a routing protocol VDTN-ToD based on DTN which uses a metric Trend of Delivery (ToD) scheme to assist in its routing and forwarding decisions. In our current work, we use this metric in order to provide better performance for DTN routing protocols Spray-And-Wait and PROPHET in VANETs. The results show that the inclusion of ToD in VANETs allows significant performance improvements and it can also be used in many other routing protocols to overcome performance issues.**

Keywords-VANET; DTN; SUMO; ToD; NS-3.

I. Introduction

The TCP/IP architecture is largely robust to deal with infrastructure networks, where a disconnection is improbable and the end-to-end path between two source/destination nodes hardly broken. This feature changes in a Mobile Ad Hoc Network (MANET) environment, where nodes are mobile and are operating in relative disconnected mode. In such conditions the TCP has its performance degraded [1]. Such a problem gets even harder when we imagine a scenario where the network is sparse and the nodes cannot mount and continuously retain an end-to-end route, which is what also happens in VANETs.

In VANETs with these problematic conditions the use of DTN architecture [2] (primarily designed for IPN routing protocols which can withstand huge delays, connections disruptions and minimizes the number of roundtrips response confirmation) is considered and proposed as a suitable alternative. DTN is also applicable in all other types of mobile networks such as cellular and wireless sensor networks. In these scenarios it is necessary to develop new protocols which know how to take advantage of the DTN paradigm. Particularly, the so-called Store-carry and Forward with random or controlled movement of mobile nodes called ferries are looked for. This allows the preamble information for connecting and routing to be in a single packet as a complete data packet, which permits the node to retain for a long time until delivery to the next participating node is successful.

Some DTN protocols are: Spray-And-Wait [3], PROPHET [4] Epidemic [5] and MaxProp [6]. However, they do not consider the specific restrictions of VANETs. Some protocols for VANETs that have been proposed based on the technique DTN are: VDTN-ToD [7], FFRDV [8], VADD [9], and GeOpps [10].

In our previous work [7], we proposed a routing protocol VDTN-ToD which uses a metric called Trend of Delivery (ToD) mechanism to assist in its routing and forwarding decisions. In this our current work, the ToD is inserted in Spray-And-Wait, MaxProp and PROPHET protocols; thus they take into account features that are specific to VANETs. The results show a considerable improvement in their performance in a VANET environment. For this, all protocols were developed for simulation in Network Simulator 3 (NS3) [11], in this article the comparisons are based on Spray-And-Wait and PROPHET protocols.

The other sections in this paper are organized as follows: Section 2 presents related work, Section 3 describes the theoretical basis with a brief overview of DTN, its architecture and routing as well as the ToD mechanism. Section 4 describes how the ToD has been incorporated into Spray-And-Wait and PROPHET protocols. Section 5 shows and describes the scenarios used together with the results from our work. Finally, in Section 6, we conclude the work and present some suggestions for future research work.

II. Related Work

A VANET environment has characteristics that hinder the existence of an end-to-end path between source and destination; therefore, DTN routing protocols have been designed for vehicular scenarios, some are reported below.

The VDTN-ToD [7] uses the metric Trend of Delivery (ToD) to assist in routing decisions to allow a particular network node to decide when it is best to keep, forward or copy a packet, taking into account improvements in the delivery rates and decrease in the message delays. The VDTN-ToD also uses a scheme of *disclosure* and *maintenance* of *location messages* based on the concept of *Adaptive Coverage Detection* (ACD), which takes into account the transmission range, to reduce the number of update messages with their location details given by the nodes.

Yu and Ko [8] proposed the VANET/DTN protocol called Fastest-Ferry Routing in DTN-enabled Vehicular Ad Hoc Networks (FFRDV). It works specifically in motor-highway scenarios. It divides the highway into blocks and within them it decides to which vehicle as a relay node it will forward the packet, based on the vehicle speed. The ToD that we incorporated into the DTN routing protocols in our work reported in this paper also uses the speed factor to assist in the routing decisions, but we go one step further taking into account the angle between the vehicle and the distance to the target node.

Zhao and Cao in [9] proposed the protocol Vehicle-Assisted Data Delivery in Vehicular Ad Hoc Networks (VADD) that uses a digital map to obtain the maximum speed, the vehicle density and intersection places. Based on this information, it uses a metric called expected delay for delivery to make routing decisions when one arrives at an intersection/junction. When it is not at an intersection it typically works like Greedy Perimeter Stateless Routing (GPSR) [12]. The VADD does not perform packet replication, but forwarding, and furthermore the target nodes are fixed in the proposed application. In our proposed scheme in the ToD protocols reported in this paper, we go one step further to ensure knowing each target's geolocation of the vehicle nodes, which are e evaluated on their mobility.

Another VANET/DTN protocol that uses metrics for its routing decisions is the Geographical Opportunistic Routing for Vehicular Networks (GeOpps) [10], which is obtained with the aid of a navigation system that makes a node to know the routes of its neighbors in it vicinity range. Since the navigation system indicates which way the neighbors will take (based on their source-to-destination route selection), suggest that GeOpps may achieve more optimum routing than in other protocols with ToD. However, this scheme may expose user security and privacy that could be used by criminals and other agencies for the wrong reasons. Protocols with ToD have the advantage of requiring less information from the VANET environment. Another proposal GeoSpray [13], which is a combination of GeOpps with Spray-And-Wait extends .

In our research work reported in this paper, the ToD mechanism has been incorporated in the Spray-And-Wait and PROPHET protocols to determine and compare their respective performances in a VANET environment.

III. Theoretical Basis

A. Delay Tolerant Networks

Initially in the 90s, IPN project was developed aimed to define the architecture for land interoperability internet with an interplanetary one. It was reported that the solutions used in IPN could also work for terrestrial networks that faced problems of disconnections and disruptions [14].

The DTN architecture uses the strategy called Store-carry and Forward, in which the first packet as a package is fully received at an intermediate node, then it stores the packet and carry (forward) it until it reaches its target destination.

The packet may be stored for hours or even days, depending on the life time set for the packet. This functionality is placed in a new layer called Layer Bundle [15], which is located below the application layer and above the transport layer.

The DTN applications generate messages of different size called bundles and they are processed, stored and forwarded in DTN nodes.

B. DTN Routing

The traditional routing protocols for networks on Earth assume that they establish an optimum end-to-end path between source and target according to some metric, such as number of nodal hops. In DTN, the concept is to establish a journey so that the bundle reaches its target by taking the maximum advantage of possible contacts (opportunity to send data) that occur with the nodes to maximize the delivery rate as quickly as possible, because there is no guarantee that a particular bundle enroute through the network will reach its destination. The protocol also has to manage the use of storage space of nodes, since in some schemes, such as epidemic routing, it can quickly fill the buffers of these devices. Another important metric is the delay metric to ensure that the protocol can deliver bundles to the target as fast as possible.

C. Trend of Delivery (ToD) Mechanism

The value of ToD is achieved through the use of fuzzy logic from soft computing that seeks to discover through the mobility of nodes how good that node is to forward or copy a packet. The ToD is based on three variables: direction ($\omega_{i,d}$), distance ($\Psi_{i,d}$) and speed ($\tau_{i,d}$), where i is the node with the message and d is the final destination of the message.

The direction indicates how close the direction from i to d is, thus a θ angle formed between the direction vector $\overrightarrow{u}$ indicating the direction of vector i and facing the recipient $\overrightarrow{v}$ is calculated, so the angle between them indicates how good or bad the value ($\omega_{i,d}$) is. The associated values are *great*, *good*, *bad* and *awful*.

For distance, four values are considered: *very close*, *close*, *far* and *very far*; each of them is achieved by the value of the transmission range of the vehicles as nodes.

In the case of the speed, four values are considered: *low*, *medium* and *high*, which indicates how fast the vehicle as a node travelling.

With the values of the variables ($\omega_{i,d}$), ($\Psi_{i,d}$) and ($\tau_{i,d}$), the value of ToD is set in seven parameter values: *maximum*

(MA), *great* (GR), *very good* (VG), *good* (GO), *bad* (BA), *very bad* (VB) and *awful* (AW).

IV. ToD Applied to Protocols

A. ToD Applied to Spray-And-Wait

The idea used to implement ToD in the Spray-And-Wait protocol to limit the number L of copies, thus the ToD is applied to assist to choose the L copies; whereas in Spray-And-Wait, these copies are spread to the first neighbors found in their immediate vicinity. The pseudo-code below shows how the decision is made:

```
/*
* Consider j as being the best
* neighbor of i, ie neighbor with
* best ToD for the bundle m chosen
* ALPHA 0.05
*/
if (node is source && L(i) > 1) {
    L(i) = L(i) / 2;
    L(j) = L(i);
    i copies the bundle m to j;
}
else if (val(ToD (i, m)) + ALPHA
  <= val(ToD(j, m))) {
    if ([(Tod(j, m)] is subset of
            [Maximum, Great, Very Good]) {
        i forwards the bundle m to j;
    }
    else if (L(i) > 1) {
        L(i) = L(i) / 2;
        i copies the bundle m to j;
    }
    else {
        i keeps the bundle m;
    }
}
else {
    i keeps the bundle m;
}
```

It can be observed that before making decisions based on ToD, we check whether the node i is the source node, in which case it always copies directly to j (where the bundle is still L $>$ 1). This decision is taken due to the possibility of having fixed source nodes in some VANET scenarios. In which case, whenever it has the opportunity it spreads the bundle to its best neighbor. We also see in the pseudo code that routing decisions are different when compared to VDTN-ToD. Since the α constant indicates the minimum difference that j must have to node i. This decision is made in order to have the best selection of neighbors, since they may show values even greater of ToDs. When the ToD node j is greater than the node i with a difference greater than or equal to alpha another check is performed, which examines whether j has a ToD that is subset of [Maximum, Great, Very Good]. When this occurs, the bundle is transferred to j, since it has a high chance of finding the destination, thus avoiding spreading unnecessary copies in the network. Moreover, the bundles are transferred even when L = 1, allowing the spread to take place successfully. If the ToD of j is not a subset of [Maximum, Great, Very Good], the bundle is copied, dividing the number of copies with the two nodes. In any other case the bundle is maintained at node i.

When a node contacts another node, each one has a list of bundles and must choose the order in which bundles should be sent, thus three mechanisms were chosen. The first works with First in First out (FIFO), in the second approach, the sequence is established based on the value of L, so the first bundles are those with the highest values of L, aiming to prioritize those bundles that were less spread. The third mechanism is the same as that used in VDTN-ToD; in this case, the bundle selection is based on its ToD.

For this work we evaluated the behavior of 4 versions for Spray-And-Wait, as follows: **Spray-And-Wait** original version using FIFO, **Spray-And-Wait V1** original version using queue approach based in the number of copies of L, **TrendOfSpray** version with trend of delivery using the same approach queue V1 and **TrendOfSpray 2** version with trend of delivery using the same approach as VDTN-ToD queue.

B. ToD Applied to PROPHET

PROPHET has its own routing strategy which is based on nodal encounter history. Hence, we applied the ToD strategy associated with the PROPHET strategy. Two approaches called PROPHETorToD and PROPHET+ToD were created.

The first works by performing an "or" between the two strategies when a bundle (chosen from the top of the queue) is ready for forwarding when it meets the sending conditions of the PROPHET protocol. If the protocol does not authorize the sending, then it goes to the strategy based on VDTN-ToD. This approach is shown in the pseudo code below:

```
/* Given the bundle m, that i
 * need to send to the destination
 * d and i have a set of n neighbors
 * P(k, d) -> Probability of node k
 * find the node d
 */

    best_neighbor =
        neighbor_with_best((P(k, d));

    if ( P(best_neighbor, d) >
            P(i, d)) {
         i copies the bundle to
          the best_neighbor;
    }
    else {
         i transfers the bundle queue
          to strategy of VDTN-ToD;
    }
```

In the second approach, the first bundle is always chosen; thus seeking the neighbor j that adds the greater delivery probability value added to ToD according to the pseudo-code below:

```
/* Given the bundle m, that i
 * need to send to the destination
 * d and i have a set of n neighbors
 */
```

```
for each k that is a neighbor of i
  j = neighbor with highest sum
    of ToD(k, m) + P(k, d)

sum_j = val(ToD(j, m)) + P(j, d);
sum_i = val(ToD(i, m)) + P(i, d);

if (sum_j > sum_i) {
    i forwards the bundle m to j;
}
else if (val(ToD(j, m)) >= val(ToD(i, m))
  || P(j, d) >= P(i, d)) {
    i copies the bundle m to j;
}
else {
    i keeps the bundle m;
}
```

V. Experiments and Results

A. Scenarios Description

For the experiments, two scenarios developed in Simulation of Urban Mobility (SUMO) [16] were prepared and the behavior of the protocols in two different scenario applications in a VANET environment was evaluated as follows.

1) Scenario 1: The first scenario is shown in Figure 1. This simpler scenario was used to evaluate a type of VANET application in which there are five nodes exchanging messages among them in the form a chats between vehicles or files exchanges.

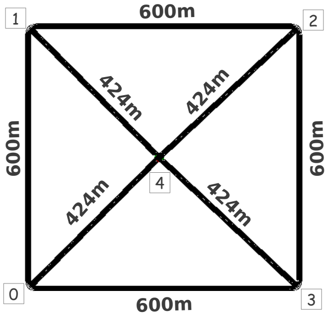


Fig. 1. Scenario 1 of simulation

TABLE I. Scenario 1 Configurations

Parameter	Configuration
Simulated Environment Area	600 x 600 m^2
Transmission Range	300 m
Maximum Speed of Nodes (Varies depending on the vehicle)	(10, 15, 20 and 25) m/s
Propagation Model	Nakagami
Model Mobility	carFollowing-Krauss (SUMO Default)
Size of Bundles	(512, 1024, 2048, 5096) bytes
Number of Generated Bundles	244
Simulation Time	300 seconds
Bundle Lifetime	200 seconds
Amount of simulations for each scenario	33
Confidence Interval	95%

All vehicles are randomly generated at the scenario edges and move randomly throughout the simulation period. The value of L for Spray-And-Wait is according to Equation 1

$$L = (N * 0.1) + 1 \tag{1}$$

where N is the number of network nodes, whose value is approximately 10% as suggested in [3]. For PROPHET values are PINIT = 0.5, γ = 0.98 and β = 0.25. Other details of the scenario are shown in Table I.

2) Scenario 2: Scenario 2 is similar to the one proposed in the previous work [7], as shown in Figure 2. In this scenario, application 4 points (0, 1, 2, 3), which are fixed DTN regions that exchange data with each other using vehicles as a data mules, is as reported in our previous work [7] (Figure 3).

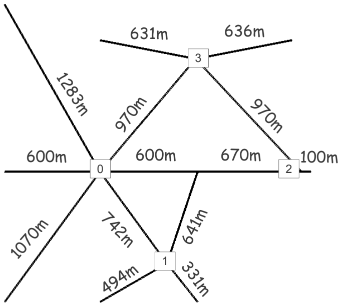


Fig. 2. Scenario 2 of simulation

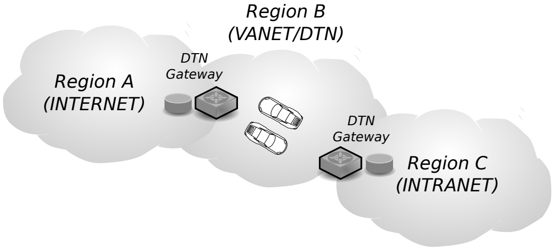


Fig. 3. Communication between remote regions

We reused this application to evaluate the protocols proposed in this work. Another detail concerning this scenario is the similarly to scenario 1: the vehicles are generated at the edges of the map and move throughout it during the simulation. Moreover, all the tracks have four lanes (two each direction).

Other information of the scenario is described in Table II.

TABLE II. Scenario 2 Configurations

Parameter	Configuration
Simulated Environment Area	1970 x 1750 m^2
Transmission Range	300 m
Maximum Speed of Nodes (Varies depending on the vehicle)	(10, 15, 20, 25) m/s
Propagation Model	Nakagami
Model Mobility	carFollowing-Krauss (Padro do SUMO)
Size of Bundles	(512, 1024, 2048, 5096) bytes
Number of Generated Bundles	471
Simulation Time	500 seconds
Bundle Lifetime	200 seconds
Amount of simulations for each scenario	33
Confidence Interval	95%

B. Metrics

Several studies suggest which metrics should be used to evaluate the performance of a DTN routing protocol. The three metrics are suggested, Delivery Rate, Average Delay, and Overhead [7] [17]. The calculations of these metrics are the same as reported in our previous work [7].

C. Analysis of ToD applied to Spray-and-Wait

We begin the analysis with the first scenario, where the VANET environment is also simpler.

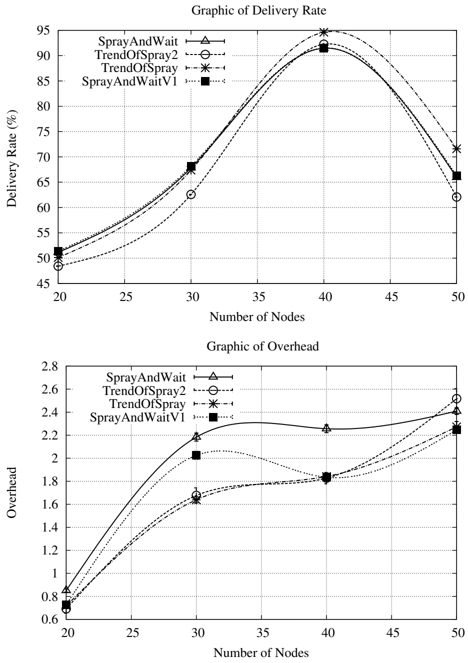


Fig. 4. *Spray-And-Wait* - Graphics of delivery rate and overhead for scenario 1

By evaluating the results in Figure 4, it can be observed that there was a slight improvement over TrendOfSpray when there was an increase in the number of vehicles, it achieved greater delivery rates, and thus keeping the overhead low. Only in once instance it achievd lower improvment to V1. Figure 5 shows a lower average delay for TrendOfSpray in all cases. These results come from a better choice of L copies that are spread associated with the strategy of the queue based on the value of L. Hence, the overhead is more controlled, since a bundle is only spread to a neighbor with a higher ToD. The bundles with these smart copies of information, tend to reach the destination faster, thus keeping the average delay lower.

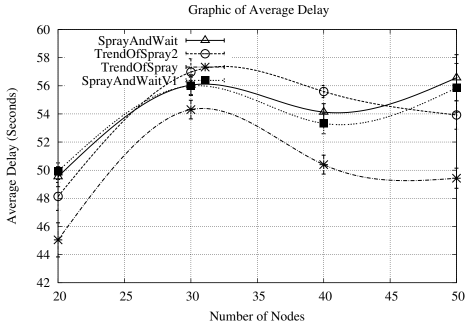


Fig. 5. *Spray-And-Wait* - Graphic of average delay for scenario 1

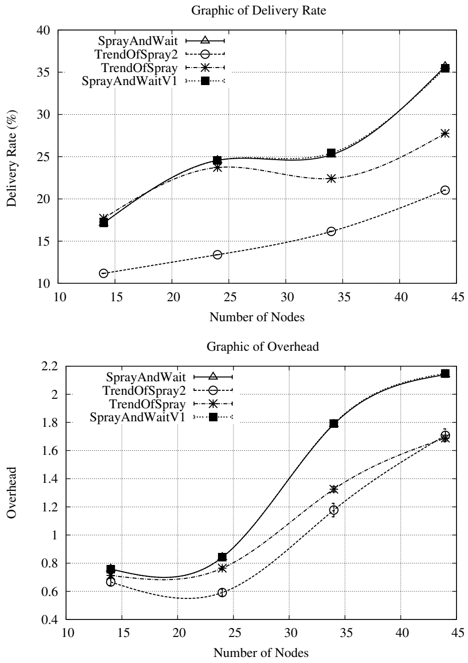


Fig. 6. *Spray-And-Wait* - Graphics of delivery rate and overhead for scenario 2

According to the results (Figures 6 and 7), it can be observed that both TrendOfSprays had delivery rates below the two versions of the Spray-And-Wait. This is due to the scenario being more complex, where the guarantee that a bundle reach its target is very low. In the case of the two TrendOfSpray approaches, several bundles are retained because the neighbor does not have a higher ToD, but it is possible that it finds its target late, since the environment is much more sparse, allowing the protocol to spread more copies speedily, hence the possibility of finding the target is greater. So, in this case the delivery rate versions of Spray-And-Wait are higher, becasue they only retain bundles when L = 1. The low overhead of TrendOfSpray approaches are reflections of bundles that are retained. Regarding the average delay, the versions of TrendOfSpray showed higher values, since they take longer to spread the bundles to network.

The TrendOfSpray (version 1) provides better results for scenario 1, keeping good delivery rates, low overhead, as well as achieving shorter delay than the two versions of Spray-And-Wait. However, in scenario 2, it did not achieve better results due to the bundles not spreading widely. A possible improvement could be to try to calibrate a better value for α; for instance, it can be 0, making the bundles to achieve more spreading in the network.

D. Analysis of ToD applied to PROPHET

From the analysis, it is important to know that PROPHET suffer more difficulties in a VANET environment of scenario

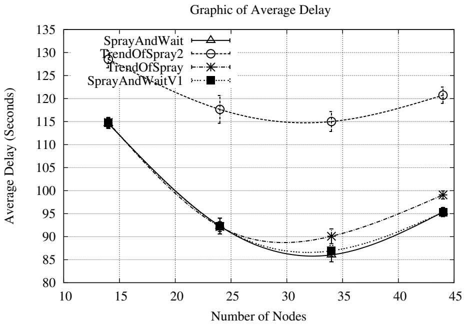


Fig. 7. *Spray-And-Wait* - Graphic of average delay for scenario 2

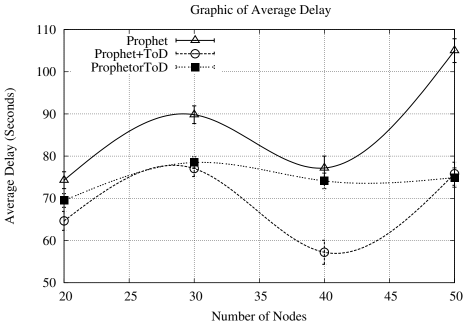


Fig. 9. PROPHET - Graphic of average delay for scenario 1

2, because it depends on a history of reencounters [7], since the bundles are forwarded only when there are reencounters or when the transitivity case occurs. This is made more difficult because both source and target nodes are fixed.

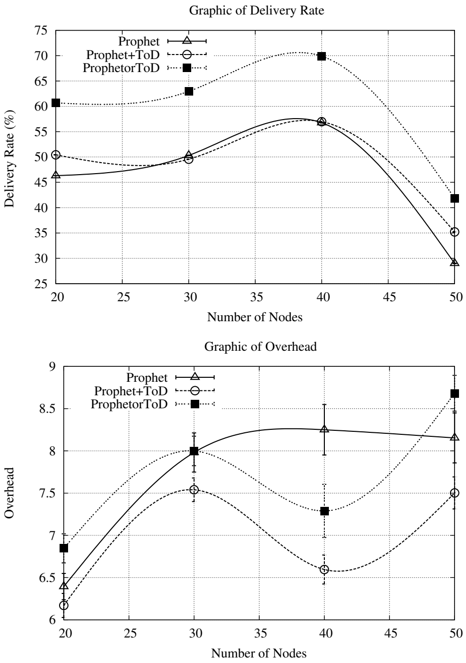


Fig. 8. PROPHET - Graphics of delivery rate and overhead for scenario 1

By observing Figure 8, it can be noticed that PROPHETorToD achieves superior delivery rates than others, with the variable overhead in relation to PROPHET (sometimes gaining, sometimes losing). The gain was due to greater probability to spread the bundles, allowing PROPHETorToD to spread the copies even if the metric PROPHET did not authorize it. in this case, with VDTN-ToD approving, with this the overhead was well controlled. For the case of PROPHET+ToD, it kept a delivery rate similarly to PROPHET, but with much lower overhead. This is due to PROPHET+ToD uses two metrics to better help their routing decisions. From Figure 9 it can be seen that compared to average delay, the PROPHET+ToD and PROPHETorToD showed a better performance since more bundles are spread using the two metrics ToD and delivery predictable. So, on the average, the bundle arrives at their target faster with greater regularity.

Observing the results shown in Figure 10, the PROPHET achieves much lower performance. In this scenario, both PROPHET and PROPHETorToD retain more bundles, making the PROPHET+ToD to achive a higher delivery rate. In the case of PROPHETorToD, it suffers from the problem of the bundles being retained longer in fixed source node (since in this case it retains the bundles due to PROPHET and VTDN-ToD not authorizing forwarding or copying), which hinders the possibility of the bundle arriving speedily at its target. With more retained bundles, the PROPHETorToD and PROPHET have lower overheads.

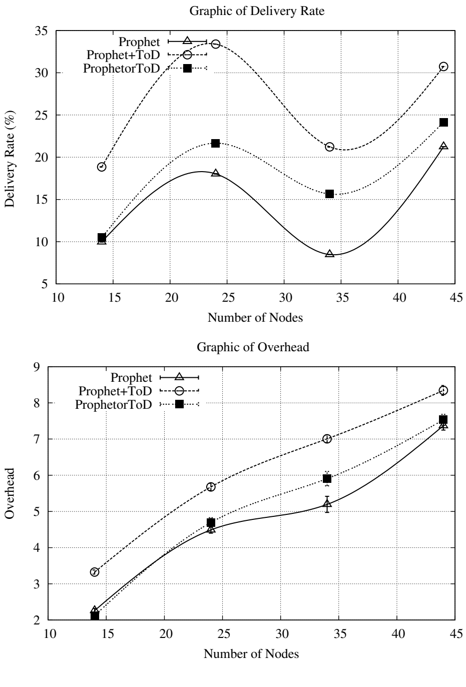


Fig. 10. PROPHET - Graphics of delivery rate and overhead for scenario 2

Referring to Figure 11, it can be seen that all of these

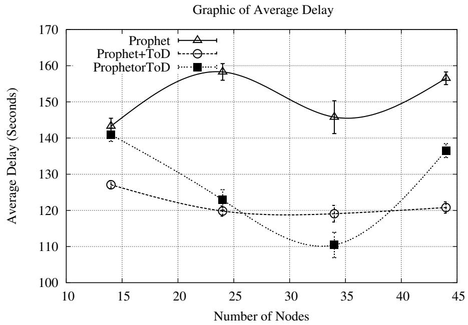


Fig. 11. PROPHET - Graphic of average delay for scenario 2

conditions of scenario 2 the PROPHET+ToD achieve a better value for average delay, since the bundles spread faster.

VI. Conclusion and Future Work

The proposed mechanism, called ToD, has succeeded in improving the performance of traditional DTN algorithms when they are applied in VANET environment. This mechanism has been also tested with the MaxProp routing protocol, and it will be reported in another follow-up paper.

The Spray-And-Wait and PROPHET protocols, using the ToD, had a significant improvement in the evaluated metrics. In the case of Spray-And-Wait, this result was expected, since it does not use any criteria for the scattering of bundles. For PROPHET, the ToD expanded the possibility of spreading the bundles, which was depended solely on historical encounters.

As future work, we consider implementing VADD and GeOpps protocols in NS3, incorporating the ToD mechanism and compare them to VDTN-ToD to perform more comprehensive evaluation.

References

[1] Z. Fu, X. Meng, and S. Lu, "How bad TCP can perform in mobile ad hoc networks," in *Computers and Communications, 2002. Proceedings. ISCC 2002. Seventh International Symposium on*. IEEE, 2002, pp. 298–303.

[2] K. Fall, "A delay-tolerant network architecture for challenged internets," in *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*. ACM, 2003, pp. 27–34.

[3] T. Spyropoulos, K. Psounis, and C. Raghavendra, "Spray and Wait: an efficient routing scheme for intermittently connected mobile networks," in *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*. ACM, 2005, pp. 252–259.

[4] A. Lindgren, A. Doria, and O. Schelén, "Probabilistic routing in intermittently connected networks," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 7, no. 3, pp. 19–20, 2003.

[5] A. Vahdat, D. Becker *et al.*, "Epidemic routing for partially connected ad hoc networks," Technical Report CS-200006, Duke University, Tech. Rep., 2000.

[6] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine, "MaxProp: Routing for vehicle-based disruption-tolerant networks." in *INFOCOM*, vol. 6, 2006, pp. 1–11.

[7] A. S. de Sousa Vieira, J. Gonçalves Filho, J. Celestino Júnior, and A. Patel, "VDTN-ToD: Routing protocol vanet/dtn based on trend of delivery," in *AICT 2013, The Ninth Advanced International Conference on Telecommunications*, 2013, pp. 135–141.

[8] D. Yu and Y. Ko, "FFRDV: fastest-ferry routing in dtn-enabled vehicular ad hoc networks," in *Advanced Communication Technology, 2009. ICACT 2009. 11th International Conference on*, vol. 2. IEEE, 2009, pp. 1410–1414.

[9] J. Zhao and G. Cao, "VADD: Vehicle-assisted data delivery in vehicular," *Vehicular Technology, IEEE Transactions on*, vol. 57, no. 3, pp. 1910–1922, 2008.

[10] I. Leontiadis and C. Mascolo, "GeOpps: Geographical opportunistic routing for vehicular networks," in *World of Wireless, Mobile and Multimedia Networks, 2007. WoWMoM 2007. IEEE International Symposium on a*. IEEE, 2007, pp. 1–6.

[11] "Network Simulator 3," access date: 28 June. 2014. [Online]. Available: www.nsnam.org

[12] B. Karp and H.-T. Kung, "GPSR: Greedy Perimeter Stateless Routing for wireless networks," in *Proceedings of the 6th annual international conference on Mobile computing and networking*. ACM, 2000, pp. 243–254.

[13] V. N. Soares, J. J. Rodrigues, and F. Farahmand, "GeoSpray: A geographic routing protocol for vehicular delay-tolerant networks," *Information Fusion*, 2011.

[14] C. T. de Oliveira, M. D. Moreira, M. G. Rubinstein, L. H. M. Costa, and O. C. M. Duarte, "Redes tolerantes a atrasos e desconexoes," *SBRC Simpósio Brasieliro de Redes de Computadores e Sistemas Distribuudos*, 2007.

[15] K. Scott and S. Burleigh, "Bundle protocol specification," 2007.

[16] D. Krajzewicz, G. Hertkorn, C. Rossel, and P. Wagner, "SUMO (Simulation of Urban MObility); An open-source traffic simulation," in *4th Middle East Symposium on Simulation and Modelling (MESM2002)*, 2002, pp. 183–187.

[17] Y. Cao and Z. Sun, "Routing in delay/disruption tolerant networks: A taxonomy, survey and challenges," *Communications Surveys & Tutorials, IEEE*, vol. 15, no. 2, pp. 654–677, 2013.

A Novel Chaos Based ASK–OOK Communication System

Branislav Jovic
Defence Technology Agency
New Zealand Defence Force
Auckland, New Zealand
b.jovic@dta.mil.nz

***Abstract*–This paper proposes the novel broadband chaos and chirp based noncoherent amplitude shift keying (ASK) – on off keying (OOK) communication systems. Their theoretical probability of error expressions are derived and confirmed with the corresponding empirical bit error rate (BER) simulations demonstrating that they match the BER performance of the traditional narrowband system. The superiority of the newly proposed broadband systems over the traditional narrowband systems is then demonstrated in the presence of narrowband interference. This is a particularly important find as the traditional narrowband ASK–OOK systems are known to be highly susceptible to interference. Finally, it is demonstrated that the interference performance of the proposed chaos based system can be further improved by optimising chaos based signals for BER using the downhill simplex algorithm.**

Keywords-Chaos; communications; ASK-OOK; interference;

I. INTRODUCTION

In chaotic communication systems a message signal directly modulates a broadband chaotic signal [1]-[3]. However, it is also possible to indirectly utilise a chaotic signal [4][5] to modulate a sinusoidal signal and thus create a constant envelope chaos based signal [5]. The chaos based signal may then be further modulated by a binary message signal to form a broadband amplitude shift keying (ASK) – on off keying (OOK) communication system as proposed in this paper. The ASK–OOK communication systems find use in radio frequency identification (RFID) devices as well as in medical applications such as endoscopy [6], among other. However, the traditional narrowband sine based ASK–OOK systems [7], where bits 0 are represented by no signal and bits 1 by a sinusoidal carrier, are highly sensitive to interference [6]. Therefore, the motivation for the use of the broadband chaos and chirp based carrier signals in place of a narrowband sinusoidal carrier signal within the ASK–OOK systems stems from the fact that broadband signals are more resistant to interference than narrowband signals.

In Section II, a novel broadband chaos based ASK–OOK system is proposed and its theoretical probability of error expression derived and confirmed by the corresponding empirical BER. Section III proposes the novel broadband chirp based ASK–OOK system and demonstrates its matching BER performance to that of the chaos based system and the traditional narrowband system. In Section IV, it is demonstrated that that the novel chaos based ASK–OOK system offers increased resistance to the narrowband interference when compared to the traditional narrowband sine based carrier and the novel broadband chirp based carrier systems. Finally, Section IV also shows that the novel broadband chaos based system can be optimised to further improve its BER in the presence of narrowband interference.

II. CHAOS BASED ASK-OOK SYSTEM

The proposed broadband chaos based ASK–OOK communication system is shown in Fig. 1 where a binary signal m modulates a continuous chaos based signal. A systematic approach of generating continuous time signals by chaos generators was proposed in [4]. The concept was used in [5] to generate the sinusoidal chaos based signals for radar applications. A chaos based signal, similar to that of [5], is generated here for the application within the novel communication system by modulating the sinusoidal carrier of (1):

$$s(t) = A\sin\left[\frac{2\pi t}{7\xi(j)}\right] \quad \text{where:} \quad [t = 0, 1 \ldots t < 7\xi(j)] \qquad (1)$$

by a chaotic signal $\xi(j)$. The chaos based signal, $s(t)$, is then modulated by a binary message signal, $m(t)$, as expressed by (2):

$$x(t) = m(t) \cdot s(t) \qquad \text{where:} \quad m(t) \in \{0, 1\} \qquad (2)$$

Finally, the modulated signal of (2), $x(t)$, is transmitted over the additive white Gaussian noise (AWGN) channel, as illustrated in Fig. 1.

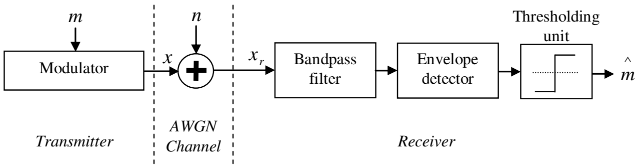


Figure 1. The noncoherent chaos based ASK-OOK wireless communication system in an AWGN channel.

ISBN: 978-1-61208-360-5

The bandpass filter of Fig. 1 filters out all but the frequencies surrounding the centre frequency of the received signal, x_r. Theoretically, the minimum bandwidth of the bandpass filter that can be implemented is equal to the bit rate $R_b = (1/T_b)$ Hertz, where T_b denotes the bit period. Accordingly, the bandwidth of the chaos based, that is, the transmitted signal, must be kept within the system bandwidth R_b.

The total fixed number of samples within the chaos based carrier signal was chosen to be 100, that is, 100 samples were chosen to represent any bit. The sample length of 100 was found to adequately represent a chaos based carrier signal of (1) and also yield a sufficiently low number of samples per bit what allowed for relatively fast Monte Carlo BER simulations. As it has not been established whether the signal $s(t)$ of (1) is chaotic, it is referred to as a chaos based rather than a chaotic signal [5]. The chaotic map $\xi(j)$ of (1), used to modulate (1), was proposed by Carroll in [5]:

$$\begin{aligned} y_{j+1}(1) &= \left[\sum_{i=2}^{N} p(i)\, y_j(i)\right] \mathrm{mod}\, 1, \\ y_{j+1}(i) &= y_j(i+1) \qquad (i = 2,3 \ldots N-1), \\ y_{j+1}(N) &= y_j(1), \\ \xi(j) &= y_j(1) + 0.5 \qquad (j = 1,2 \ldots M), \end{aligned} \tag{3}$$

where $p(i)$ are the system parameters. The largest Lyapunov exponent of $\xi(j)$ was determined to be 0.097 bits/iteration [5] thus proving that the proposed system of (3) is chaotic. For a system to be chaotic, its largest Lyapunov exponent must be greater than 0 [8].

For every chaotic sample of (3), there is a certain higher number of chaos based signal samples of (1) [5]. The number of chaos based signal samples per chaotic value is higher for higher chaotic signal values and lower for lower values. The reason for this can be observed by examining (1) from which it can be seen that higher $\xi(j)$ chaotic values produce lower frequency components and more samples per each sinusoidal cycle of the chaos based signal. Opposite is true for lower $\xi(j)$ chaotic values which produce higher frequency components and less samples per sinusoidal cycle of the chaos based signal. Therefore, in contrast to a linear frequency modulated (LFM) chirp signal [9] whose frequency changes linearly, frequency components of a chaos based signal change in a chaotic fashion.

The empirical distributions of the decision variables of the newly proposed broadband chaos based ASK–OOK system of Fig. 1 are plotted in Fig. 2 at the bit energy to noise power spectral density (*Eb/No*) ratio of 10 dB, demonstrating the Rayleigh and Rician distribution of bits 0 and 1, respectively.

As the energy of a narrowband sinusoidal signal $(A^2T/2)$ [8][10] is independent of its frequency and only depends on its amplitude and signal duration, it can also be shown that the energy expression of the sinusoidal chaos based signal of (1) is governed by the same expression. To demonstrate this, (1) is first rewritten in the form of (4):

$$s(t) = \sin[2\pi\,\Delta f\; t] \tag{4}$$

where $\Delta f = 1/[7\,\xi(j)]$ denotes the changing frequency with chaotic values. The chaos based signal energy is then obtained by squaring and integrating (4) over the signal duration T:

$$\begin{aligned} E &= \int_0^T (A\sin[2\pi\,\Delta f\; t])^2\, dt \\ &= \frac{A^2}{2}\int_0^T dt - \frac{A^2}{2}\int_0^T (\cos[4\pi\,\Delta f\; t])\, dt \\ &= \frac{A^2 T}{2} \end{aligned} \tag{5}$$

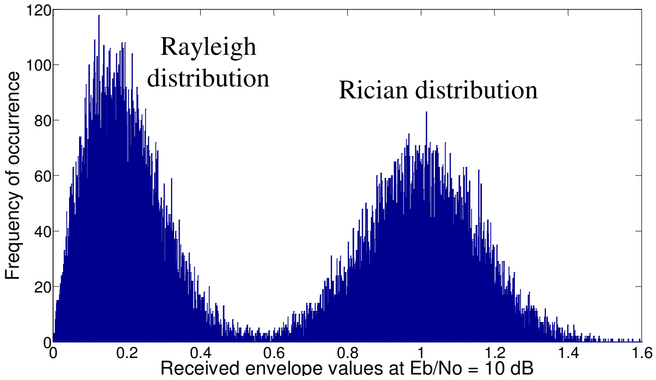


Figure 2. Histogram of the received envelope values at the *Eb/No* ratio of 10 dB for the novel chaos based ASK–OOK system.

By setting the signal duration T of (5) equal to the bit period *Tb*, bit 1 energy equal to $A^2T_b/2$ is obtained. Furthermore, as the bits 0 are represented by no signal their energy is zero, resulting in an average energy per bit of $E_b = A^2T_b/4$ [7]. Therefore, the chaos based signal energy is identical to that of a narrowband sinusoidal signal [7][10].

As the distribution of the decision variables and the energy of the chaos based signal match those of the traditional narrowband sine based ASK-OOK system [7] it is expected that the probability of error of the two systems will be identical. The probability of error (P_e) is obtained by considering the distribution of the decision variables sampled from the envelope at the input to the thresholding unit of Fig. 1. Accordingly, the probability of error that a bit 1 is received when a bit 0 was transmitted is obtained by integrating the Rayleigh probability density function representing bits 0 for the values when the decision threshold $\alpha = A/2$ is exceeded:

$$P_{e_0}(\alpha) = \int_\alpha^\infty \frac{r}{\sigma^2} e^{-r^2/(2\sigma^2)}\, dr = e^{-\alpha^2/(2\sigma^2)} \tag{6}$$

where σ^2 denotes the variance (power) of the AWGN at the input of the envelope detector, A the amplitude of the chaos based carrier and r the random variable. The noise power at the input to the envelope detector is bandwidth limited by the filter of bandwidth $B_p = R_b$ and is thus expressed as: $\sigma^2 = \frac{N_o}{2} 2B_p = N_o R_b$. Similarly, the probability that a bit 0 is received when a bit 1 was transmitted is obtained by integrating the Rician probability density function representing bits 1 for the values when the decision threshold α is not exceeded:

$$P_{e_1}(\alpha) = \int_0^{\alpha} \frac{r}{\sigma^2} e^{-(r^2+A^2)/(2\sigma^2)} I_o\left(\frac{rA}{\sigma^2}\right) dr = 1 - Q_M\left(\frac{A}{\sigma}, \frac{\alpha}{\sigma}\right) \qquad (7)$$

where I_o denotes the modified Bessel function of the first kind of zero order. The integral term of (7) was put into the form of the Marcum Q function [11]: $Q_M(a,\beta) = \int_\beta^\infty x\, e^{(-(x^2+a^2)/2)} I_o(ax) dx$ by making the definition: $x = r/\sigma$, so that $dx = dr/\sigma$ and substituting those into the integral term to obtain the final result of (7).

The probability of error for equiprobable bits 0 and 1 is then determined by finding the average value of (6) and (7), to obtain the exact P_e expression for the noncoherent chaos based ASK-OOK system:

$$P_e(\alpha) = \frac{1}{2}\left(e^{-\alpha^2/(2\sigma^2)} + 1 - Q_M\left(\frac{A}{\sigma}, \frac{\alpha}{\sigma}\right)\right) \qquad (8)$$

Equation (8) is plotted in Fig. 3a alongside the corresponding empirical BER curve obtained via the Monte Carlo simulation of the chaos based ASK–OOK system of Fig. 1. It can be observed that the two plots exhibit a perfect match and thus prove the result of (8). Furthermore, as expected, the result of (8) matches the result for the probability of error of the traditional narrowband ASK-OOK system whose theoretical probability of error expression was derived in [7]: $P_e(\alpha) = 0.5[e^{-A^2/(8\sigma^2)} + Q(A/(2\sigma))]$, where Q denotes the Gaussian Q function. However, due to a number of simplifying assumptions made in the derivation of [7] the final expression is somewhat inaccurate for high levels of noise, as demonstrated in Fig. 3b where it is plotted against the more complex but exact result of (8) and the narrowband empirical BER.

As opposed to indirectly using a chaotic signal to modulate the sinusoidal carrier within the chaos based ASK–OOK system, it is also possible to use a chaotic signal directly to represent a bit 1 and no signal to represent a bit 0. Such a system, termed chaotic on off keying (COOK), was proposed and investigated in [12]-[14]. However, unlike sinusoidal signals, chaotic signals generally do not have a constant envelope making a COOK system potentially more difficult and costly to implement.

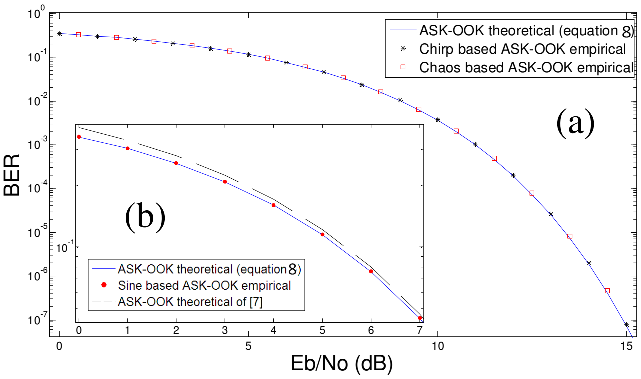


Figure 3. (a). Theoretical and empirical BER curves of the newly proposed broadband systems. (b). Margin of error within the theoretical BER of [7].

III. Chirp based ASK-OOK system

By replacing the chaos based carrier signal by an LFM chirp signal, a novel broadband ASK–OOK system is proposed here. Its novelty over the chirp based spread spectrum OOK system of [6] is in the envelope detector receiver architecture, which is simpler and more cost effective than the power detector architecture of [6] that was empirically investigated in the human body channel. The power detector architecture involves squaring and integration of the received signal, as opposed to the envelope detector architecture which simply samples the constant envelope, resulting in a lower implementation cost. As for the chaos based system, it was found here that the decision variables of the proposed system are also of Rayleigh and Rician distributions. Furthermore, as the energy of an LFM chirp signal is also equal to $A^2T/2$ [9], resulting in an average energy of a bit of $E_b = A^2 T_b / 4$, it is readily verifiable by repeating the derivation presented in (6) to (8) that the probability of error of the novel chirp based system is also governed by (8). The result is confirmed in Fig. 3a by an exact match between the empirical BER curve and the theoretical result of (8).

IV. BER performance and optimisation in the AWGN channel with narrowband interference

This section presents the BER performance of the narrowband sinusoidal and the two novel broadband systems when subjected to the sinusoidal narrowband interference. The novel chaos based system is then optimised to demonstrate a further improvement in the BER performance.

A. BER performance of the ASK-OOK systems

The BER results for the signal to interference ratios (SIR) of 20, 16, 12, 10, 8 and 6 dB are plotted in Fig. 4 from where it can be observed that although the narrowband and the proposed broadband ASK–OOK systems exhibit matching performance in the AWGN channel only, the broadband systems exhibit superior performance at all levels of interference.

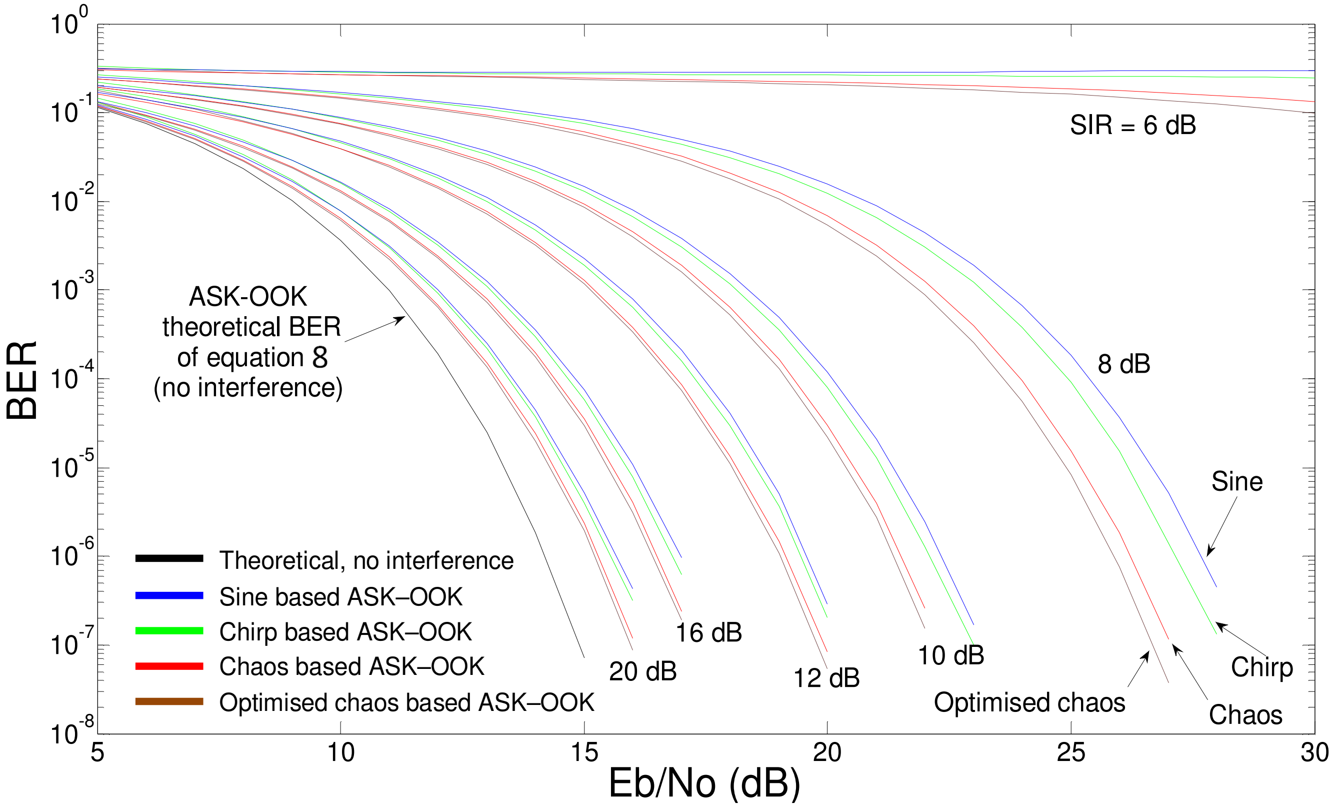


Figure 4. Empirical BER of the ASK–OOK systems for varying SIR levels.

The newly proposed broadband chirp based ASK–OOK system exhibits an increasing BER improvement over the narrowband system with the decreasing SIR. A further similar, but more notable, improvement over the narrowband sinusoidal and the broadband chirp based system is exhibited by the novel broadband chaos based system. For instance, at the SIR of 12 dB, and the BER level of 10^{-4}, an improvement of approximately 0.5 and 0.75 dB can be observed in favour of the chaos based system when compared to the chirp and narrowband sinusoidal systems, respectively. The improvement increases to approximately 1 and 1.5 dB at the SIR level of 8 dB. The bandwidth of the chirp carrier signal was kept the same as that of the chaos based carrier signal while the narrowband interference tone was kept near the centre frequency and within the system bandwidth R_b. For the narrowband sinusoidal ASK–OOK system, the interference tone was generated at the carrier frequency.

B. Optimisation of the novel chaos based ASK-OOK system

In [5], chaos based signals were used for radar applications where broadband chaos based radar pulses were optimised for detection of specific targets. In contrast to an LFM chirp signal where only a single unique signal can be generated in a given bandwidth, a chaotic system can theoretically produce an infinite number of chaotic signals and therefore also chaos based signals by varying its parameters or initial conditions. By producing a large number of different chaos based signals, Carroll showed [5] that it is possible to find a particular chaos based signal that yields a high signal reflection from one arbitrary radar target and a low signal reflection from some other arbitrary radar target. This was achieved by optimising chaos based signals for a given radar target, or a set of targets, whose reflection characteristics are known in advance. The optimisation of the chaos based radar model was achieved by employing the downhill simplex or Nelder–Mead algorithm [15]. Downhill simplex algorithm determines the minimum of a given equation, termed minimisation equation. In case of the chaos based radar optimisation problem the minimisation equation of [5] was based on minimising the reflection of one radar target with respect to another.

Taking the lead of [5], it is now demonstrated that the BER performance of the proposed chaos based ASK–OOK system of Section II can be further improved by optimising chaos based signals for a given level of narrowband interference. As opposed to minimising reflections from radar targets, the downhill simplex algorithm was used here to minimise the BER of the communication system. Optimisations were performed for each SIR level investigated at an *Eb/No* ratio corresponding to the BER level of 10^{-4} of a non optimised chaos based ASK–OOK system. Once an optimisation for a given SIR level was complete, the system was evaluated with the determined

ISBN: 978-1-61208-360-5

optimum chaos based carrier for a range of *Eb/No* ratios and the empirical BER curve obtained. The optimisations were started off with an initial random set of parameters that produce a non optimised chaotic signal of (3) and thus also a chaos based signal of (1). Upon evaluating the BER for the initial set of parameters, the downhill simplex algorithm varied the parameters and re-evaluated the BER. The process was repeated a large number of times until a smallest local minimum (optimum) BER value was determined. The optimisations were all nonlinear resulting in many local minima (as opposed to a definite global minimum), of which some had similar magnitudes. Nonetheless, the astringency of the algorithm may be improved by starting the algorithm execution with a favourable initial set of chaotic parameters. However, again, as the process is nonlinear, there is no rule on how to choose initial chaotic parameters that may enhance astringency.

An advantage of the downhill simplex algorithm over some other optimisation algorithms is that it does not require derivatives [15] what reduces its complexity. Other algorithms may also be employed to try to further improve the BER performance [16][15]. However, this is beyond the scope of this paper and will not be investigated here.

The optimised empirical BER curves for the varying SIR levels are plotted in Fig. 4 on the same set of axes as the non optimised curves. From the plots it can be observed that the improvement in BER performance of the optimised over non optimised chaos based ASK–OOK system ranges from less than 0.1 to approximately 0.5 dB at different SIR levels. Therefore, although the observed improvements are less significant than those of subsection IV-A, they further widen the improvement margin over the traditional narrowband system and demonstrate that the proposed chaos based ASK–OOK system's BER performance can be further improved through an optimisation process.

V. Conclusions

In this paper, two novel broadband chaos and chirp based ASK–OOK communication systems were proposed. Their theoretical probability of error expression was derived and confirmed by an exact match with the corresponding empirical BER simulations. It was shown that both of the proposed novel broadband ASK–OOK systems exhibit superior BER performance to the narrowband system in the presence of narrowband interference, while also exhibiting a matching performance in the AWGN channel only. The novel chaos based system was shown to exhibit the best BER performance with interference in the system, providing an improvement of up to 1.5 dB. Such a find is of particular importance for the ASK–OOK systems, which are traditionally known to offer little resistance to interference. Furthermore, it was demonstrated that the newly proposed chaos based ASK–OOK system can be optimised using the downhill simplex algorithm to further improve its BER performance by up to 0.5 dB in the presence of narrowband tone interference.

References

[1] B. Jovic, Synchronization Techniques for Chaotic Communication Systems, Springer-Verlag, Berlin Heidelberg, 2011.

[2] U. Parlitz and S. Ergezinger, "Robust communication based on chaotic spreading sequences", Phys. Lett. A, vol. 188, num. 2, May 1994, pp. 146-150.

[3] B. Jovic and C.P. Unsworth, "Fast synchronisation of chaotic maps for secure chaotic communications", IET Letters, vol. 46, num. 1, January 2010, pp. 49-50.

[4] A. L. Baranovski and W. Schwarz, "Statistical Analysis and Design of Continuous-Discrete Chaos Generators", IEICE Trans. Fundamentals, vol. E82-A, num. 9, September 1999, pp. 1762-1768.

[5] T. L. Carroll, "Optimizing chaos-based signals for complex radar targets", Chaos, vol. 17, num. 3, September 2007, pp. 033103-1–033103-10.

[6] M. Jeon, K. Kim, and J. Lee, "Interference Reduction Modulation Based on Chirp Spread Spectrum for Capsule Endoscopy", IEEE Workshop on Sig. Proc. Syst., Quebec City, October 2012, pp. 91-96.

[7] L. W. Couch, Digital and Analog Communication Systems, Prentice Hall International, Upper Saddle River, 5th edn. 1996, p. 479.

[8] J. C. Sprott, Chaos and Time-Series Analysis, Oxford University Press, Oxford, pp. 104, 2003.

[9] A. J. Berni and W. D. Gregg, "On The Utility of Chirp Modulation for Digital Signalling", IEEE Trans. on Com, vol. 21, num. 6, June 1973, pp. 748-751.

[10] S. Haykin, Communication Systems, Wiley, New York, 4th edn., pp. 349-350, 2001.

[11] J. S. Lee and L. E. Miller, CDMA Systems Engineering Handbook, Artech House Publishers, Boston, pp. 124, 1998.

[12] G. Kolumban, H. Dedieu, J. Schweizer, J. Ennitis, and B. Vizvki, "Performance evaluation and comparison of chaos communication schemes", Proc. of the 4th International Workshop on Nonlinear Dynamics of Electronic Systems, Seville, Spain, June 1996, pp. 105-110.

[13] G. Kolumban, M. P. Kennedy, and G. Kis, "Performance improvement of chaotic communication systems", Proc. of the 13th European Conference on Circuit Theory and Design, Budapest, Hungary, September 1997, pp. 284-289.

[14] G. Kolumban, M. P. Kennedy, Z. Jako, and G. Kis, "Chaotic Communications with Correlator Receivers: Theory and Performance Limits", Proc. of the IEEE, May 2002, pp. 711-732.

[15] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, Numerical recipes in C, the art of scientific computing, Cambridge University press, New York, 2nd edn. 1994, pp. 408-412, 1988.

[16] M. E. Homer, S. J. Hogan, M. di Bernardo, and C. Williams, "The Importance of Choosing Attractors for Optimizing Chaotic Communications", IEEE Trans. on Circuits and Systems – II: Express Briefs, vol. 51, num. 10, October 2004, pp.511-516.

Hybrid QoS Based Routing for IEEE 802.16j Mesh Infrastructure

Hajer Bargaoui, Nader Mbarek, Olivier Togni
LE2I Laboratory, University of Burgundy
Dijon, France
e-mails : {Hajer.Bargaoui, Nader.Mbarek, Olivier.Togni}@u-bourgogne.fr

Mounir Frikha
MEDIATRON Laboratory, High School of Communication of Tunis (SUP'COM)
Tunis, Tunisia
e-mail : m.frikha@supcom.rnu.tn

***Abstract*—With the growth of wireless networks, Wireless Mesh Network (WMN) has appeared as an emerging key solution for broadband Internet access with a low-cost deployment. Moreover, providing QoS guarantees for real-time and streaming applications such as VoIP (Voice over IP) and VoD (Video on Demand) is a challenging issue in such environment. In this paper, we propose a hybrid wireless mesh architecture to provide mesh clients with Internet access while guaranteeing QoS. It is formed by an IEEE 802.16j based infrastructure and several IEEE 802.11s based client domains. A clustering algorithm is developed to enhance scalability issues within the mesh infrastructure and a novel protocol called Hybrid QoS Mesh Routing (HQMR) is specified in order to provide QoS requirements. The HQMR protocol is deployed within the IEEE 802.16j infrastructure and it is composed of two routing sub-protocols: a reactive routing protocol for intra-infrastructure communications and a proactive QoS-based multi-tree routing protocol for communications with external networks. The proposed architecture provides real-time and streaming applications with QoS guarantee in mesh environment thanks to a clustering algorithm and a QoS-based routing protocol.**

Keywords—Wireless Mesh Network; QoS routing; IEEE 802.16j; HQMR.

I. INTRODUCTION

Recently, wireless mesh networks have received increased attention from researchers and industrial environments. They have emerged as a key wireless technology for numerous applications such as broadband home networking, community and neighborhood networks, enterprise networking, etc., [1][2]. Besides, they are a promising solution to provide last-mile connectivity to the Internet for fixed and/or mobile users in zones where wired networks deployment is difficult, thanks to its various qualities such as self-organizing and self-configuring abilities.

One major challenge for wireless mesh networks is to provide QoS support. Since deployments of WMNs continue to grow, providing Quality of Service for real-time and streaming applications, such as VoIP and VoD, is an important task. Moreover, establishing paths with the highest performance is a challenging issue for routing protocols on wireless mesh networks in order to satisfy applications' requirements.

However, the different research works proposing routing solutions on wireless mesh networks rely simply on adapting protocols originally designed for mobile ad hoc networks and adding a little support for QoS. In this paper, we propose a hybrid QoS based routing protocol, called Hybrid QoS Mesh Routing (HQMR) that exploits more efficiently the particular topology of a wireless mesh network, based on a hybrid wireless mesh architecture. The proposed wireless mesh architecture is formed by an IEEE 802.16j based infrastructure and different IEEE 802.11s based client domains. Furthermore, in order to solve scalability issues and reduce efficiently the network's load, a clustering algorithm is proposed for the IEEE 802.16j infrastructure of our global wireless mesh architecture. HQMR is then deployed on the IEEE 802.16j infrastructure to ensure routing functionalities. It is a hybrid protocol adopting a reactive routing sub-protocol for intra-infrastructure communications and a proactive multipath tree-based routing sub-protocol for inter-infrastructure communications, where the mesh gateway is considered as a root.

The remainder of this paper is organized as follows. In Section II, we present some related works. Section III introduces the architecture of our framework. Then, we define in Section IV, the proposed HQMR routing protocol. Section V defines two usage scenarios of HQMR to illustrate its processing. Finally, Section VI concludes the paper.

II. RELATED WORK

A. IEEE 802.16j Standard

IEEE 802.16j task group was officially established in March 2006 and their work was published in 2009. The IEEE 802.16j standard [3], is an amendment to the IEEE 802.16e [4] standard in order to introduce Mobile Multi-hop Relay (MMR) specifications where traffic between a Multi-Relay Base Station (MR-BS) and a Subscriber Station (SS) can be relayed through nodes named Relay Stations (RS). The number of hops between MR-BS and SS is not defined but it must only contain RS nodes. In fact, IEEE 802.16j has defined two different relay modes: transparent mode and non-transparent mode. In transparent mode, the RS is used to improve the network capacity. It does not forward any signaling frame. It relays only data traffic. The non-transparent mode is usually used to extend the network coverage. The RS nodes in this mode are able to generate their own signaling frame or forward those provided by the MR-BS depending on the scheduling mechanism.

B. QoS Routing

QoS provisioning is an important issue for wireless mesh networks since they are typically used for providing broadband wireless Internet access to a large number of users

ISBN: 978-1-61208-360-5

and networks. To meet applications' QoS requirements, different QoS routing protocols were proposed for wireless mesh networks.

Wireless Mesh Routing (WMR) [5] is a QoS solution for wireless mesh LAN networks. It provides QoS guarantees in terms of minimum bandwidth and maximum end-to-end delay. These two parameters are verified jointly with the route discovery process. The value of the node's available bandwidth, is estimated thanks to the bandwidth already in use by the considered node and by its neighboring nodes. Then, the end-to-end delay is estimated by using the round trip delay method [6]. Kon et al. [7] improve the WMR protocol by proposing a novel end-to-end packet delay estimation mechanism with a stability-aware routing policy. The delay estimation is based on packets named DUMMY-RREP, which have the same size, priority and data rate as real data traffic.

Some other works include the QoS verification in the route discovery phase. For example, QoS AODV (QAODV) [8] integrates a new metric for IEEE 802.11 mesh networks, composed of bandwidth, delay, hop count and load ratio. In the same way, Rate-Aware AODV (R-AODV) [9] uses minimum network layer transmission time as a performance metric in multi-rate WiFi mesh networks. Mesh Admission control and qos Routing with Interference Awareness (MARIA) [10] is another QoS aware routing protocol for wireless mesh networks. It is a reactive protocol incorporating an interference model in the route discovery process. This protocol uses a conflict graph model to characterize both inter and intra-flow interference.

Thus, there are few research works for QoS based routing for IEEE 802.16j wireless networks. Hence, through our proposed protocol HQMR, we intend to provide QoS provisioning routing functionalities within an IEEE 802.16j architecture.

C. *Clustering*

Clustering concept was introduced to organize large wireless multi-hop networks into groups named clusters. Every cluster is coordinated by a cluster-head to achieve basic network performances, even with mobility and limited energy resources. The different clustering algorithms differ mainly in the method used for the election of the cluster-heads: Lowest-ID heuristic [11], Highest-degree heuristic [12] and node-Weight heuristic [13].

Combining clustering algorithms with routing protocols offers better performances within the network layer, by reducing the amount of control messages propagated inside the network since the exchange is limited within a cluster; and by minimizing the size of routing tables at each node since it stores only the information of its cluster.

Zone Routing Protocol (ZRP) [14] is a cluster-based routing protocol for ad hoc networks that uses different routing sub-protocols for inter and intra-clusters communications. Within a cluster zone, a proactive component is used to maintain up-to-date routing tables. Routes outside the routing zone are explored with a reactive component combined with a border-casting concept. Singh et al. [15] propose a hierarchical cluster based routing protocol for wireless mesh networks in which the mesh gateway is the highest level node. Similarly, the research work in [16] defines a multi-level clustering approach with a reactive routing protocol for wireless mesh networks, in order to reduce the load on the mesh gateway.

For its benefits, we adopt this concept of cluster based routing for our HQMR protocol to solve scalability issues and to offer better routing performances within the IEEE 802.16j infrastructure.

III. PROPOSED GLOBAL HYBRID WIRELESS MESH ARCHITECTURE

For our framework, we adopt a hybrid wireless mesh network architecture, combining two different technologies. It is formed by a non-transparent IEEE 802.16j-based infrastructure and IEEE 802.11s-based client domains (Fig. 1). A hybrid QoS based routing protocol (HQMR) is also proposed within the 802.16j-based wireless mesh infrastructure.

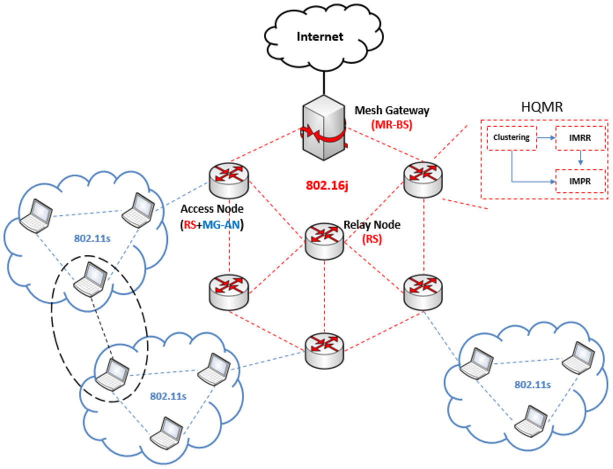


Figure 1. Global hybrid wireless mesh architecture

A. *IEEE 802.16j-based mesh infrastructure domain*

For the wireless mesh infrastructure, we use the non-transparent relay mode of the IEEE 802.16j technology to ensure a better coverage. Then, in order to organize the functionalities of each node, we define three types of nodes within the mesh infrastructure: the Mesh Gateway (MG), the Relay Nodes (RN) and the Access Nodes (AN). The MG is the intermediate node between the Internet cloud and the wireless mesh infrastructure. It helps forwarding clients requests to the Internet network. The RNs are the nodes located in the core of the mesh infrastructure to ensure forwarding traffic flows from a node to another inside it. Last, we consider the nodes located in the border of the infrastructure, as Access Nodes (AN). They provide interconnection between the mesh infrastructure and the client domains. Thus, compared to the topology of an IEEE 802.16j network, our MG and RN nodes have, respectively, the same functionalities as the MR-BS node and the RS nodes. In fact, the AN nodes may be considered as bridge nodes playing both the role of a relay node in the IEEE 802.16j infrastructure and the role of a gateway in the IEEE 802.11s area. Thus, they are equipped with two radio interfaces: one is operating with the Wimax technology [4] and another with WiFi technology [17].

At each relay node of the wireless mesh infrastructure (including the ANs and the MG), the proposed routing protocol (HQMR) must be implemented with our clustering algorithm to reduce mainly the size of the routing tables. The different blocks of HQMR will be described in Section IV.

B. IEEE 802.11s based mesh client domain

The client domains are formed by a set of 802.11s [18] MP (Mesh Point) which are interconnected to each other forming the mesh topology and by a gateway node that we called Mesh-Gateway Access Node (MG-AN). The MG-ANs have the functionality of the 802.11s MPP (Mesh Portal Point) implemented in the access node (AN) of our mesh infrastructure. This way, to connect to the Internet cloud, the mesh clients forward, first, their traffic to their own gateway (i.e., MG-AN), for accessing the mesh infrastructure. Then, the MG-AN forwards directly the received traffic from its mesh clients to its own gateway (i.e., MG).

IV. HYBRID QOS MESH ROUTING

HQMR, the proposed protocol, is used to ensure routing functionalities within the IEEE 802.16j infrastructure of our global wireless mesh architecture. It is a hybrid QoS-based routing protocol composed of two different routing blocks. The first routing sub-protocol Intra-Mesh infrastructure Reactive Routing (IMRR) is designed to forward communications within the infrastructure in a reactive manner, while the second routing block Inter-Mesh infrastructure Proactive Routing (IMPR) is deployed to forward communications to the external networks, particularly to the Internet network. The second routing sub-protocol is a tree-based multipath routing protocol, with the Mesh Gateway as a root of the routing tree.

Moreover, in order to improve the performance of our routing protocol, we adopt the concept of clustering to divide the topology of the infrastructure into a set of groups.

In this section, we present the algorithm adopted for the clusters elaboration within the wireless mesh infrastructure and we introduce the two routing sub-protocols of HQMR. Before that, we define the mechanism used to provide the needed information about each node's neighbors and we specify the different QoS parameters and their estimation method to guaranty the QoS based routing characteristic of our proposed HQMR protocol.

A. Neighborhood Maintenance

Neighborhood information is very important for our protocol in order to provide the local topology (node's different neighbors), the necessary information for our clustering algorithm and the available QoS toward each neighbor. To maintain this information, every node in the network is required to send out periodically a Hello message (Table I), announcing its existence and its cluster information such as its state in the cluster, its calculated weight parameter used for cluster-head election, its CH's IP address (ID-CH) and its used bandwidth parameter. By receiving the Hello message from the different neighbors, each node updates its Neighbor Table (Table II), which is used to store for each neighbor its IP address (ID), all the needed information for clusters formation (Weight, State, ID-CH) and the available QoS parameters.

TABLE I. HELLO MESSAGE

ID	Weight	State	ID-CH	Used Bandwidth

TABLE II. NEIGHBOR TABLE (NT)

ID	Weight	State	ID-CH	QoS Metric

B. QoS Routing Metrics

The purpose of our routing protocol is to find paths, which can satisfy the QoS requirements of real-time flows. The set of QoS requirements includes the bandwidth, the delay and the jitter parameters.

1) Available Bandwidth metric

To estimate the available bandwidth, each node considers the used bandwidth by its flows and the consumption of its neighbors announced in the Hello messages (1).

$$B(v) = B - \sum_{v' \in N(v)} B_{used}(v') \qquad (1)$$

where B(v) is the estimated available bandwidth by a node v, B is the total Bandwidth, B_{used} is the bandwidth used by a node and N(v) in the neighborhood of the node v.

Then, the bandwidth parameter of the entire path is determined as the minimum bandwidth estimated at each node toward the destination.

2) Delay Metric

This metric estimation is based on measuring the round trip delay time (RTT) [6] of the Hello messages, which represents the time between initiating a Hello message and receiving a response. The delay metric of a path is the sum of its links delay metric.

3) Jitter Metric

The jitter metric defines the delay metric variation. It is estimated by calculating the mean of the differences between the RTT values for a specific period. Besides, the Jitter of a path is calculated by summing the Jitter of each link.

C. Clusters formation algorithm

Our clustering algorithm is a variant of the LID-based clustering algorithm [11] combined with the use of the weight concept developed on the Weighted Clustering Algorithm (WCA) [13] for the election of cluster-heads. Thus, a cluster is formed by the node with the lowest weight and all its neighbors. The same procedure is repeated among the remaining nodes, until each node is assigned to a cluster. Inter-clusters connectivity is maintained by defining some Gateway-nodes (sub-section 3), named Cluster Gateway (C-Gw) and Distributed Gateway (D-Gw). Moreover, in our adapted algorithm, we have opted for one-hop clusters to reduce the load of control messages within a cluster and to ensure a line of sight between the different cluster-heads and gateway nodes, which is an important characteristic for the deployment of our second routing sub-protocol IMPR (section E). An example of a clustered wireless mesh infrastructure is illustrated in Fig. 2.

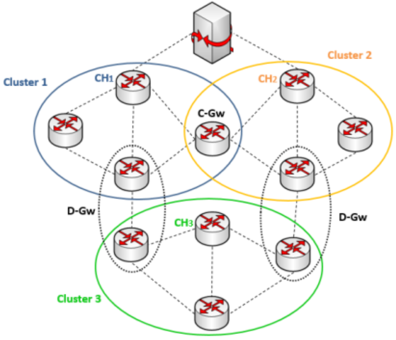


Figure 2. Clustered architecture of the wireless mesh infrastructure

Our clustering algorithm is composed of three main functions, which are presented in the following sub-sections: weight calculation, cluster-head election and clusters elaboration process.

1) *Weight Calculation*

In our algorithm, the weight assigned to each node is based on the WCA algorithm [13]. The latter takes into account the degree (neighbors' number), the transmission power, the mobility and the battery power of each node. It optimizes the degree of each cluster-head by choosing an optimal number M of nodes per cluster (M is a pre-defined threshold). This restriction aims that the cluster-head would be able to support ideally the nodes within its cluster.

However, given the stability of the nodes within our wireless mesh infrastructure, we are only interested in the first two parameters used to calculate the weight of WCA to find the optimal number of nodes within the transmission range and to estimate the transmission power toward the neighbors of a node. In addition, since most of the traffic is oriented to the Mesh Gateway, a third parameter is used in our weight calculation to take into account the power transmission of the node toward the Mesh Gateway. By this way, the cluster-head will be elected among the nearest nodes to the Mesh Gateway. Thus, the weight is calculated according to (2)-(6):

$$Wv = a \cdot \Delta v + b \cdot Dv + c \cdot DPv \tag{2}$$

where a, b and c are the weighing factors so that a+b+c=1 and Wv is the weight of a node v.

$$dv = \left|N(v)\right| = \sum_{v' \in V, v' \neq v} (dist(v,v')\langle tx_{range}) \tag{3}$$

where V is the neighborhood of a node v.

$$\Delta v = \left|dv - M\right| \tag{4}$$

$$Dv = \sum_{v' \in N(v)} dist(v,v') \tag{5}$$

$$DPv = dist(v, MG) \tag{6}$$

Equation (4) represents the degree-difference for a node v to compare its number of neighbors (3) to the optimal number of nodes that a CH may coordinate efficiently. The transmission power toward the neighbors is estimated in (5) by computing the sum of the distances with all its neighbors. Samely, the third parameter namely the transmission power toward the Mesh Gateway is calculated in (6).

2) *Cluster-head Election*

Initially, all the nodes are in the initial state that is the "Undecided" state and with a weight equal to zero. Thanks to the periodic exchange of Hello messages, the Neighbor Table (Table II.) will be updated with the last calculated value of weight (W) for each neighbor. Each node waits for a period T_e before starting the selection of the cluster-heads, so that all the nodes have updated their NT. After this period, the node with the lowest W among its neighbors broadcasts a Hello message, as illustrated in Fig. 3.

```
If $W_i$ = min (NT [weight]) then
        $S_i$ = CH
        ID-CH$_i$ = ID$_i$
        Broadcasts Hello (ID$_i$, W$_i$, S$_i$, ID-CH$_i$, B$_{used}$)
End If
```

Figure 3. Cluster-head election algorithm

3) *Clusters elaboration process*

The division of the network on a set of clusters is based on the exchange of Hello Messages between each node and its neighbors. Fig. 4 illustrates the algorithm of the clusters elaboration.

```
On receiving a Hello message:
If (Hello [State] = CH) then {
  If ID-CH$_i$ = null then {
          S$_i$ = CM
          ID-CH$_i$ = Hello [ID-CH]
          Update (NT)
          Broadcast Hello (ID$_i$, W$_i$, S$_i$, ID-CH$_i$, B$_{used}$) }
  Else {
          If (Hello [ID] < ID-CH$_i$) then
            G = ID-CH$_i$
            ID-CH$_i$ = Hello [ID]}
          End If
          Unicast [ID-CH$_i$, GW-D (ID$_i$, W$_i$, C-Gw, ID-CH$_i$, G, null)] }
  End If }
Else if (Hello [State] = CM) then {
   If (S$_i$ = CM) then {
          If (ID-CH$_i$ = Hello [ID-CH]) then
            Update (NT)
          End If
          Unicast [ID-CH$_i$, GW-D (ID$_i$, W$_i$, D-Gw, ID-CH$_i$, Hello [ID- CH], Hello[ID])]
   }
   Else
          Update (NT)
   End If }
Else {
   If (ID-CH$_i$ = Hello [ID-CH]) then Update (NT)
   Else {
      Update (NT)
      Update (NCHT)
      }
   End If }
End If
```

Figure 4. Clusters elaboration algorithm

According to our algorithm, we distinguish five possible states of a node within a cluster. Besides, it is important to notice that the clustering algorithm is executed on each node of the infrastructure except the Mesh Gateway. The latter has its own state MG as Mesh Gateway. For the rest of nodes, we have the following states:

- **Undecided**: it is the initial state indicating that the node does not yet belong to any cluster.
- **Cluster Member** (CM): it is a node, which belongs already to a cluster.
- **Cluster Head** (CH): it is the node with the lowest weight and it is the cluster's manager.
- **Cluster Gateway** (C-Gw): it is a node in direct vision with two different cluster heads at the same time. It acts as a bridge between the two clusters.
- **Distributed Gateway** (D-Gw): it is a CM that has a neighbor belonging to another cluster. D-Gw ensures the communications between two disjoint clusters.

These different states with the different necessary transition conditions are decribed in a FSM diagram (Fig. 5).

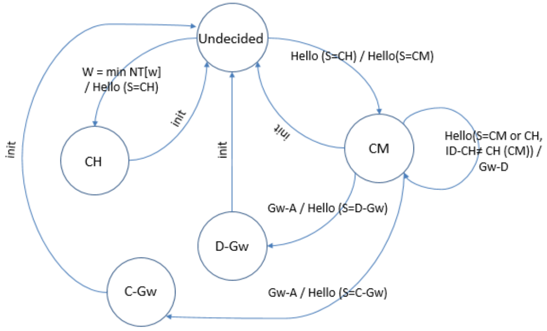

Undecided
Hello (S=CH) / Hello(S=CM)
W = min NT[w] / Hello (S=CH)
init
init
init
init
CH
CM
Hello(S=CM or CH, ID-CH≠ CH (CM)) / Gw-D
Gw-A / Hello (S=D-Gw)
D-Gw
C-Gw
Gw-A / Hello (S=C-Gw)

Figure 5. FSM of a node participating in the clustering algorithm

A node becomes a CM node when it receives a Hello message for the first time from a CH node. This node may change its state to a gateway node to ensure interconnection between two clusters. It may become a C-Gw when receiving a Hello message from another cluster-head. It sends a GW-D (Declare) message (Table V.) to its cluster-head without changing state. By receiving this message, the cluster-head consults its Neighbor CH Table (NCHT) (Table IV) in which it keeps the neighbor cluster-heads and its corresponding gateways and responds with a GW-A (Accept) or GW-R (Refuse) message. When a node is accepted as C-Gw, it changes its state to a C-GW (it is no more a CM) and updates its Gateway Table (Table III), in which it keeps its type as gateway and the two interconnected cluster-heads.

TABLE III. GATEWAY TABLE (GWT)

Type-Gw	ID-CH1	ID-CH2	ID_D-Gw

TABLE IV. NEIGHBOR CLUSTER-HEAD TABLE (NCHT)

Neighbor ID-CH	Gw-ID	Type-Gw

TABLE V. GW-D MESSAGE

ID	Weight	Type-Gw	ID-CH	Neighbor ID-CH

A CM node may also become a D-Gw when receiving a Hello message from a CM belonging to another cluster, as illustrated in the MSC diagram in Fig. 6.

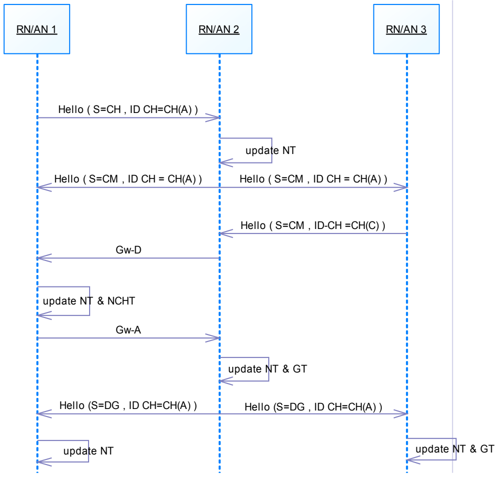

RN/AN 1
RN/AN 2
RN/AN 3
Hello (S=CH , ID CH=CH(A))
update NT
Hello (S=CM , ID CH = CH(A))
Hello (S=CM , ID CH = CH(A))
Hello (S=CM , ID-CH =CH(C))
Gw-D
update NT & NCHT
Gw-A
update NT & GT
Hello (S=DG , ID CH=CH(A))
Hello (S=DG , ID CH=CH(A))
update NT
update NT & GT

Figure 6. MSC of D-Gw selection scenario

D. Intra-infrastruture Routing (IMRR)

Intra-Mesh Infrastructure Reactive Routing (IMRR) is the reactive routing sub-protocol of our proposed HQMR protocol. It is used to find routes in order to forward information between two nodes located within the infrastructure. It ensures QoS based routing for nodes belonging to a same cluster as well for those located in different clusters. Furthermore, the proposed IMRR sub-protocol is an adaptation of AODV routing protocol [19] to take into account the clustering approach and the QoS verification in route discovery process.

1) IMRR operation

Fig. 7 illustrates the algorithm of IMRR operation. A node S starts directly to forward data if a valid route to D exists in its routing table or if D is one of its neighbors, with verified QoS. Otherwise, S launches the route discovery process.

```
When a node S wants to transmit data to a node D:
S verifies its routing table
If a valid route with requested QoS exists then  Forward (data, D)
Else {verify (NT)
    If D exists and QoS verified then Forward (data, D)
    Else send (RREQ, CH)
    End If}
End If
On receiving a RREQ message:
If QoS verified then {
    If it is the destination then Send (RREP, S)
    Else {update (RREQ); update (RT)
        If it is a CH then {Verify (NT)
            If D exists then Unicast (RREQ, D)
            Else multicast (RREQ)
            End If}
        Else multicast (RREQ)
        End If}
    End If
Else discard (RREQ)
End If
```

Figure 7. IMRR operation algorithm

The received RREQ message is either forwarded directly to the destination or forwarded to the multicast group formed by the different CHs, C-Gws, D-Gws and the MG. The use of the multicast group limits the broadcast of the RREQ messages, which helps reducing the load of the network.

We distinguish two main cases: the intra-cluster routing and the inter-clusters routing. For the first case, a node communicates with another within its cluster either in direct manner or through its cluster-head. An example of the second case is illustrated by a MSC (Message Sequence Chart) [20] in Fig. 8.

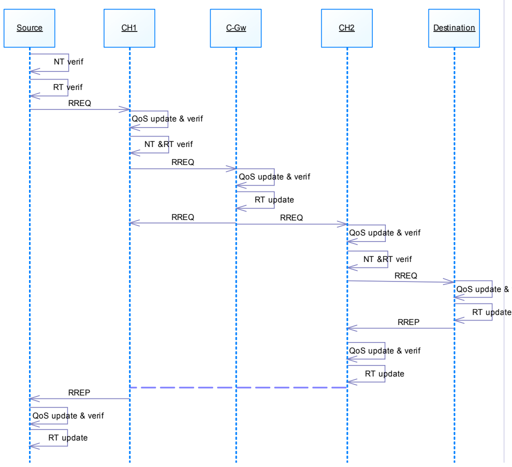


Figure 8. MSC for inter-clusters IMRR routing

Two nodes from different clusters may communicate with each other only through a route formed by CHs and/or Gws and/or the Mesh Gateway.

2) Route Discovery Process

Like AODV protocol, IMRR uses RREQ message for route discovery (Table VI). However, the RREQ message used by our IMRR routing protocol introduces specific QoS fields to enable QoS based routing. Each intermediate node proceeds to a QoS verification before forwarding the request (7).

$(B_{off}>=B_{req}$ or $B=null)$ and $(D_{off}>=D_{req}$ or $D=null)$ and $(J_{off}>=J_{req}$ or $J=null)$(7)

where B is the bandwidth, D is the delay and J is the Jitter.

TABLE VI. RREQ MESSAGE

Src IP address	Dest IP address	Broadcast ID	Path	QoS Metric request-ed	QoS Metric offered	ID msg

In Fig. 9, we illustrate the processing of a RREQ message at each node. Unlike AODV protocol, only the destination node is able to respond to a RREQ message, so that it would have the entire path's estimated QoS to compare it properly to the requested one. Moreover, the duplicate RREQ messages are not rejected. Instead, we send as much as possible of RREQ messages to the destination to guarantee the discovery of the best path. In order to avoid an infinite loop of a message, each node verifies first if its address already exists in the Path field or not. Then, we introduce a new parameter called "ID msg" to distinguish the duplicate messages at a node. This parameter is updated at each intermediate node for each RREQ message received (duplicated or not). Then, the reverse route is created within the routing table (Table VII), by taking into consideration this parameter, so that it would be used later for the RREP message forward.

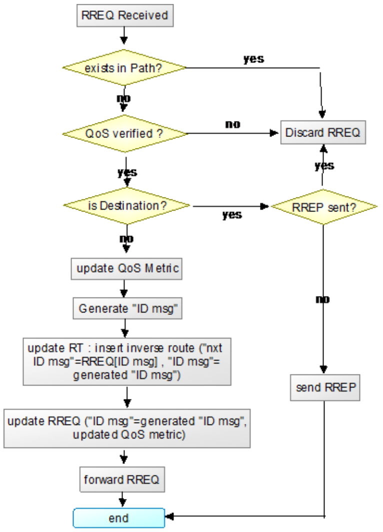


Figure 9. RREQ processing

TABLE VII. IMRR ROUTING TABLE

Dest IP address	Next Hop	Lifetime	QoS Metric offered	ID msg	Nxt ID msg

3) Route Replay Process

In order to establish a route toward the source node, the destination responds with a RREP message (Table VIII) to the first RREQ received verifying the requested QoS parameters and rejects the following RREQ messages. The processing of a RREP message at each intermediate node is illustrated by a flowchart in Fig. 10.

A mesh node determines the next hop thanks to the "ID msg" parameter. It updates then the routing table with the direct route and the "ID msg" with the "Nxt ID msg" of the routing table before forwarding the RREP message.

TABLE VIII. RREP MESSAGE

Src IP address	Dest IP address	Lifetime	QoS Metric requested	QoS Metric offered	ID msg

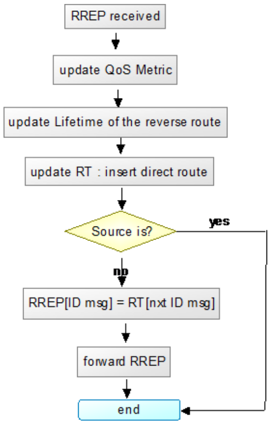

RREP received
update QoS Metric
update Lifetime of the reverse route
update RT : insert direct route
Source is?
yes
no
RREP[ID msg] = RT[nxt ID msg]
forward RREP
end

Figure 10. RREP processing

E. *Inter-infrastructure Routing (IMPR)*

Inter-infrastructure Mesh Proactive Routing (IMPR) is the second routing sub-protocol of HQMR, designed to ensure communications toward external networks, especially Internet network. Since most of the traffic goes through the Mesh Gateway to provide Internet services, we opted for a proactive tree based routing protocol, having the Mesh Gateway as a root and the different CHs and C-Gw and/or D-Gw as children. It is important to notice that the different cluster members would not participate in the trees construction process.

In addition, to provide QoS guarantees for real-time flows, IMPR deploys a multi-path routing concept to define three different routes, partially node-disjoint, between each child and the root. These routes would be used to construct three partially disjoint routing trees within the IEEE 802.16j wireless mesh infrastructure, in such a way that each tree is used to forward a specific type of traffic. To this end, we define for our protocol three service classes, namely interactive real-time applications class, Streaming applications class and Best Effort class. The first class is more sensitive to delay and jitter variations, the second one is more sensitive to jitter variation and the last class is more exigent in terms of loss ratio. In other words, IMPR allows the construction of three partially disjoint trees with a common root: Real Time, Streaming and Best Effort Trees.

1) Root Announcement process

The root (i.e., MG) broadcasts a RANN (Route Announcement) message to all its neighbors to announce its presence. This message is considered only by the CHs and the Gws. It is rejected by all the CM nodes. On receiving a RANN message (Table IX.), each intermediate node stores the Path parameter in its route cache and updates it next by adding its address. It updates also the QoS Metric and proceeds to the forward of the updated RANN message to its multicast group formed by the CHs, the Gws and the MG. In order to keep as many routes as possible, duplicated RANN messages are not rejected. Instead, to avoid an infinite loop of a message, each node verifies first if its address already exists in the Path field or not. In fact, each node keeps the entire path received through the RANN message in its route cache in order to be able to verify later the disjunction of two paths.

TABLE IX. RANN MESSAGE

Root IP address	Path	QoS Metric

2) Routing trees construction

Each node waits for a certain time Ts before starting the routing trees construction process, in order to store the maximum of paths. Firstly, using the routes selection algorithm (Fig. 11), each node selects a route for the Real Time Tree. This route is validated as one of the tree branches by an exchange of PREQ and PREP messages with the root. Once the PREP received from the root, each node removes the chosen path from its route cache and starts the construction process of the second routing tree in the same manner. Then, the mechanism is repeated for the third routing tree. In fact, the exchange of PREQ/PREP messages performed for routes validation is used to ensure that each intermediate node of a path is using the same path toward the root, so that each node has no more than a single branch toward the root of a tree.

3) Routes Selection Algorithm

This algorithm is described in Fig. 11. The idea is to select at each node a potential path for each routing tree, satisfying the requirements of the defined service classes. For the first path corresponding to the Real Time Tree, we choose the best in terms of delay and jitter with satisfying bandwidth metric. The second one should be partially disjoint from the first one to reduce congestion issues, with good values of the jitter QoS parameter. Lastly, from the remaining paths, we select the best in terms of disjunction over the other paths.

Some nodes may not be able to select three different paths. Thus, for the case where a node has only selected two paths, the first one would be used to forward the highest priority traffic, while the second one would be shared between the two other service classes . If only one path is present at a node, we adopt the default QoS mechanism of IEEE 802.16j to share it between the three service classes.

```
P ← set of stored paths ; Disj ← number of common nodes between paths
HC: Hop Count ; $w_i$ : QoS parameters' weight
If $tree_i = 1$ then {
    $A = \{P\}_{(D<Dmax \text{ and } J<Jmax)}$
    If $A \neq \Phi$ then $P_1 = min_{HC}\ \{max_{Bw}\ A\}$
    Else {
        $B = \{P\}_{(D<Dmax)}$
        If $B \neq \Phi$ then{Calculate $L=w_1*rank_{desc}Bw + w_2*rank_{asc}J$ for each path in B; $P_1 = min_L\ B$}
        Else{Calculate $L=w_1*rank_{desc}Bw+w_2*rank_{asc}D+w_3*rank_{asc}J$ for each Path in P; $P_1 = min_L\ P$}
        End If }
    End If }
End If
If $tree_i=2$ then {
    $P = P\backslash\{P_1\}$ ; $A = \{P\}_{(J<Jmax)}$
    If $A \neq \Phi$ then {
        Calculate $L=a*rank_{desc}\ Bw+b*rank_{asc}\ Disj+c*rank_{asc}J$ for each Path in A
        $P_2 = min_L\ A$}
    Else {
        Calculate $L =w_1*rank_{desc}\ Bw + w_2*rank_{asc}\ J + w_3*rank_{asc}\ Disj$ for each Path in P
        $P_2 = min_L\ P$ }
    End If }
End If
If $tree_i=3$ then {
        $P = P\backslash\{P_2\}$ ; $P_3 = min_{HC}\ \{min_{Disj}\ P\}$ }
End If
```

Figure 11. IMPR routes selection algorithm

4) Path Request Process

By executing the route selection algorithm, a node selects a path for its ith routing tree and sends a PREQ message (Table X). Each intermediate node compares its chosen path for its ith routing tree to the path carried by the PREQ message. If the next hop in the two paths is different, the node either modifies its entire path or updates the path in the PREQ message, as presented in the Flowchart in Fig. 12. Then, the intermediate node updates its routing table (Table XI) with both the direct route (toward the root) and the reverse route (toward the source) and forwards the PREQ message to the next hop.

TABLE X. PREQ MESSAGE

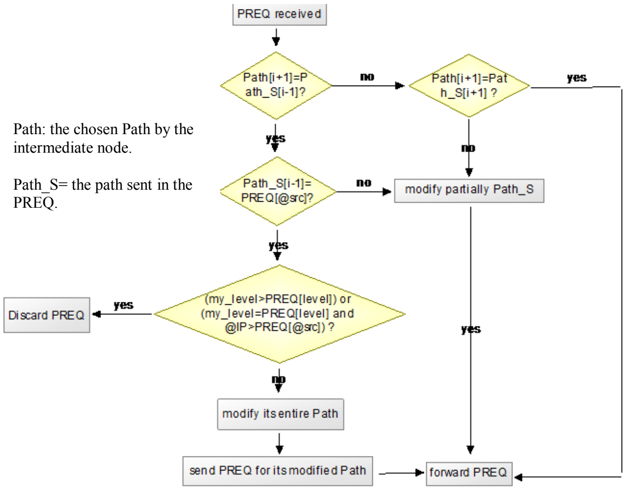
Src IP address	Dest IP address	Path	ID-Path	Level[a]

[a] Level : the level of a node in the Real Time tree

ISBN: 978-1-61208-360-5

TABLE XI. IMPR ROUTING TABLE

Dest IP address	Next Hop	ID-Path



Path: the chosen Path by the intermediate node.

Path_S= the path sent in the PREQ.

Figure 12. Flowchart of PREQ process

5) *Path Replay Process*

On receiving the PREQ message, the root updates its routing table and sends a PREP (Table XII) message to its child.

TABLE XII. PREP MESSAGE

Dest IP address	Path	ID-Path

Each intermediate node adds its address to the Path parameter of the PREP message and forwards it to the destination. Once the destination receives the PREP message, it updates its routing table and its chosen path for the routing tree if it is different from the Path parameter in the PREP message. Then, it removes it from its route cache to begin the selection of a route for the next tree.

V. HQMR USAGE SCENARIOS

In this section, we present two different usage scenarios of our HQMR protocol, describing how a path is selected to reach a destination within or outside the mesh infrastructure.

A. *Intra-infrastructure Routing Usage Scenario*

This scenario describes how to determine a QoS verified path between two nodes from different clusters for a VoIP application between two mesh clients of our architecture. To this end, the reactive routing bloc, named IMRR would be used and a RREQ message is generated for route discovery process. In Fig. 13, we illustrate the RREQ process through each intermediate node by comparing the offered QoS to the requested one (B_{req}=56Kb/s, D_{req}=150ms and J_{req}=20ms).

The first RREQ received by D (<2, 155, 19>) does not satisfy the requested delay parameter. Thus, this message is discarded and D waits for another RREQ messages. Since the second message received (RREQ2) verifies the different QoS parameters (<2, 145, 13>), a RREP message is unicasted to the source node.

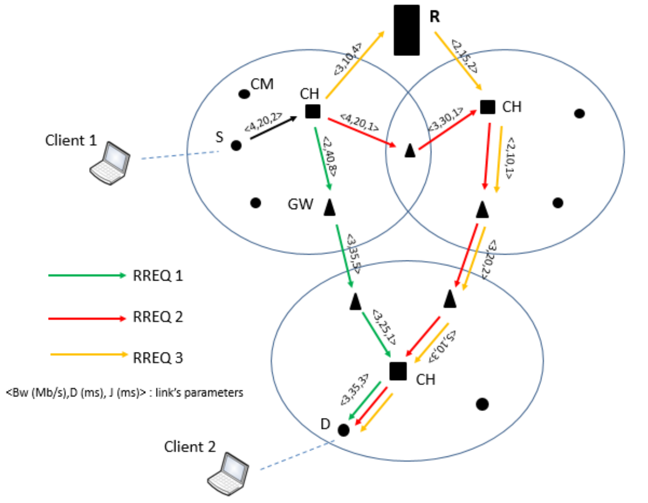


Figure 13. Intra-infrastructure usage scenario

Then, regarding the third RREQ message received, it would be discarded since a RREP message has been already sent back. By this way, the route discovered by RREQ2 would be used to forward the traffic of the VoIP application between the two mesh clients.

B. *Inter-infrastructure Routing Usage scenario*

For communications with the Internet network, the proactive routing protocol IMPR of HQMR protocol is used. In this scenario, we describe how to forward a VoD application traffic from a streaming video server in the Internet. To this end, three QoS based routing trees are constructed. Fig. 14 shows an example of clustered topology over which we have built the three QoS based routing trees.

For the first routing tree, by executing the route selection algorithm (D_{max}=150 ms, J_{max}=20 ms), we chose the paths with satisfying delay and Jitter parameters. For example, in the case of the node B, we have four paths towards the root satisfying the delay and jitter parameters: B-A-R: <4,30,11> ; B-C-R: <3,70,6> ; B-D-A-R: <3,95,16> ; B-E-C-R: <2,70,9>. Then, the path with the highest Bandwidth is selected: B-A-R. This process is repeated at each node of the topology and a PREQ message is sent for route validation.

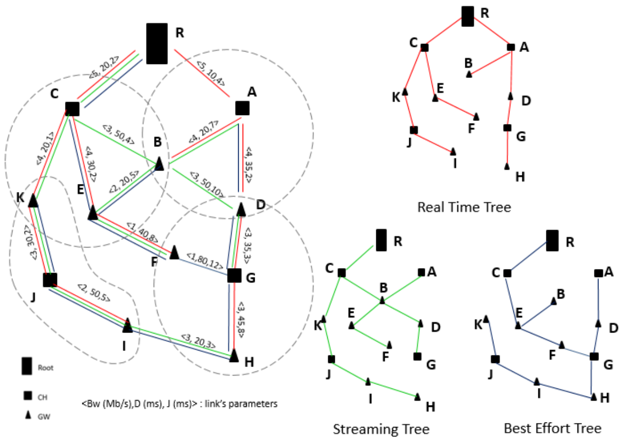


Figure 14. The QoS based routing Trees

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-360-5

Similarly, the path at each node for the second routing tree is selected according to the routes selection algorithm. For example, at the node E, we have two paths verifying the jitter parameter: E-B-C-R: <2, 90, 11>; E-B-A-R: <2, 50, 16>. The second path is the one selected by E since it offers better disjunction with the path selected for the first tree. However, the PREQ of this message would be changed at the node B. In fact, the nodes B and E have the same level parameter but node B has a greater IP address than E. Thus, the path in the PREQ sent by the node E would be changed (to E-B-C-R) to correspond to the route chosen by the node B: B-C-R.

By receiving the PREP message for the selected route, each node starts the selection of its third route that is the most disjoint route to the two first selected paths with a minimum of hops. For example, the node I according to these conditions chooses the path I-H-G-D-A-R. However, its PREQ at the node G would be changed (see the flowchart in Fig. 12). The selected path by the node I would be modified partially (I-H-G-F-E-C-R) to correspond to the one selected by the node G. Then, the set of paths selected at each node forms the routing tree for the third service class.

Regarding our usage scenario, and in order to forward the VoD traffic, the second routing tree would be used since this application is considered as an application of the Streaming service class.

VI. CONCLUSION

In this paper, we presented our proposed hybrid wireless mesh architecture composed of two different domains: an IEEE 802.16j-based infrastructure domain and several IEEE 802.11s based client domains. Then, we have specified the HQMR protocol for ensuring routing functionalities within the 802.16j infrastructure of our global architecture. It is a hybrid QoS based routing protocol formed by a reactive routing sub-protocol for a clustered infrastructure and a proactive multipath tree based routing sub-protocol for communications toward Internet network. Two usage scenarios are presented to show the importance of HQMR in order to provide real time and streaming applications with QoS guarantee in wireless mesh networks.

As a future work, we are working on the HQMR performance evaluation within our global wireless mesh architecture as well as comparison with other protocols.

ACKNOWLEDGMENT

This work has been funded by the Regional Council of Burgundy, France.

REFERENCES

[1] I. F. Akyildiz, X. Wang, and W. Wang, “Wireless mesh networks: a survey”, Comput. Netw., march. 2005, vol. 47, no 4, pp. 445-487.

[2] I. F. Akyildiz and X. Wang, “A survey on wireless mesh networks”, IEEE Commun. Mag., 2005, vol. 43, no 9, pp. S23-S30.

[3] “IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Broadband Wireless Access Systems Amendment 1: Multihop Relay Specification”, IEEE Std 80216j-2009 Amend. IEEE Std 80216-2009, 2009, pp. 1-290.

[4] “IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1”, IEEE Std 80216e-2005 IEEE Std 80216-2004Cor 1-2005 Amend. Corrigendum IEEE Std 80216-2004, 2006, pp. 01-822.

[5] Q. Xue and A. Ganz, “QoS Routing for Mesh-Based Wireless LANs”, Int. J. Wirel. Inf. Netw., 2002, vol. 9, n^o 3, pp. 179-190.

[6] R. Draves, J. Padhye, and B. Zill, “Comparison of routing metrics for static multi-hop wireless networks”, New York, NY, USA, 2004, pp. 133–144.

[7] V. Kone, S. Das, B. Y. Zhao, and H. Zheng, “QUORUM: quality of service in wireless mesh networks”, Mob Netw Appl, Dec. 2007, vol. 12, n^o 5, pp. 358–369.

[8] L. Liu, L. Zhu, L. Lin, and Q. Wu, “Improvement of AODV Routing Protocol with QoS Support in Wireless Mesh Networks”, Phys. Procedia, 2012, vol. 25, pp. 1133-1140.

[9] Y. Zhang, Y. Wei, M. Song, and J. Song, “R-AODV: Rate aware routing protocol for WiFi mesh networks”, IET International Conference on Wireless, Mobile and Multimedia Networks, 2006, pp. 1-4.

[10] X. Cheng, P. Mohapatra, S. Lee, and S. Banerjee, “MARIA: Interference-Aware Admission Control and QoS Routing in Wireless Mesh Networks”, in IEEE International Conference on Communications, ICC ’08, 2008, pp. 2865-2870.

[11] C. R. Lin and M. Gerla, “Adaptive clustering for mobile wireless networks”, IEEE J. Sel. Areas Commun., 1997, vol. 15, n^o 7, pp. 1265-1275.

[12] P. Krishna, N. H. Vaidya, M. Chatterjee, and D. K. Pradhan, “A cluster-based approach for routing in dynamic networks”, SIGCOMM Comput Commun Rev, April 1997, vol. 27, n^o 2, pp. 49–64.

[13] M. Chatterjee, S. K. Das, and D. Turgut, “WCA: A Weighted Clustering Algorithm for Mobile Ad Hoc Networks”, Clust. Comput., April 2002, vol. 5, n^o 2, pp. 193-204.

[14] Z. Haas, M. Pearlman, and P. Samar, “The Zone Routing Protocol (ZRP) for Ad Hoc Networks”, http://tools.ietf.org/html/draft-ietf-manet-zone-zrp-04 [retrieved: May,2014].

[15] M. Singh, S. G. Lee, T. W. Kit, and L. J. Huy, “Cluster-based routing scheme for Wireless Mesh Networks”, 13th International Conference on Advanced Communication Technology (ICACT), 2011, pp. 335-338.

[16] D. Kaushal, A. G. Niteshkumar, K. B. Prasann, and V. Agarwal, “Hierarchical Cluster Based Routing for Wireless Mesh Networks Using Group Head”, International Conference on Computing Sciences (ICCS), 2012, pp. 163-167.

[17] “IEEE Draft Standard for Information Technology-Telecommunications and information exchange between systems-Local and Metropolitan networks-Amendment 8: IEEE 802.11 Wireless Network Management", IEEE P80211vD160, 2010, pp. 1-428.

[18] “IEEE Standard for Information Technology–Telecommunications and information exchange between systems–Local and metropolitan area networks–Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 10: Mesh Networking”, 2011, pp. 1-372.

[19] “Ad hoc On-Demand Distance Vector Routing (AODV)”. http://www.ietf.org/rfc/rfc3561.txt [retrieved: May,2014].

[20] D. Harel and P. S. Thiagarajan, "Message Sequence Charts", in UML for Real, Ed. Springer US, 2003, pp. 77-105.

A Proposal for Path Loss Prediction in Urban Environments using Support Vector Regression

Robson D. A. Timoteo, Daniel C. Cunha
CISG - Communication and Info. Systems Research Group
Centro de Informática - UFPE
Recife - PE - Brazil
rdat@cin.ufpe.br, dcunha@cin.ufpe.br

George D. C. Cavalcanti
VIISAR - Vision and Artificial Intelligence Research Group
Centro de Informática - UFPE
Recife - PE - Brazil
gdcc@cin.ufpe.br

Abstract—**In the last few years, the mobile data traffic has grown exponentially making evident the importance of wireless networks. To ensure an acceptable level of quality of service for users in a wireless data network, network designers rely on signal propagation path loss models. To provide adaptability, the use of machine learning techniques has been considered to predict characteristics of the wireless channel. In this work, we propose a method for predicting path loss in an urban outdoor environment using support vector regression. Simulation results indicate that, depending on the employed kernel and its parameters, the performance obtained using support vector regression is similar and with lower computational complexity to that obtained by a multilayer perceptron neural network.**

Keywords*—*wireless networks, propagation models, machine learning, nonlinear regression.

I. INTRODUCTION

Today, being connected is crucially important. High-speed internet access via mobile handsets is the most likely way of achieving digital inclusion [1]. In this context, the importance of wireless and mobile networks grows every day and their demand is also becoming larger even faster.

To ensure an acceptable level of quality of service for users in a wireless data network, network designers rely on signal propagation path loss models. Radio wave propagation models are a series of mathematical calculation developed to predict path characteristics and losses in a given environment [2]. For example, propagation models have traditionally focused on predicting the average received signal strength at a given distance from the transmitter, plus the variability in the signal intensity near a particular location area. Thus, propagation models are mathematical tools used by engineers and scientists to plan and optimize wireless network systems.

Given the problem context, many researchers have turned their attention to the domain of machine learning (ML) [3]. The goal of this class of algorithms is to automatically learn the properties of the environment and to adapt their behavior quickly and easily to them. Artificial neural networks (ANN) are a typical example of a ML algorithm, inspired by the biological neural networks of the brain [4]. In recent time, multilayer perceptron (MLP)-ANN have been shown to successfully perform path loss in rural, urban, suburban and indoor environments [5], [6], [7]. However, a drawback in using MLP-ANN is the required training time to process data, considering the numerous neurons in each layer of the neural network. To handle it, other ML techniques can be used, such as support vector machine (SVM). The main advantages of using SVM are the absence of local minima, the sparseness of the solution and the capacity control obtained by optimising the margin [8]. Aside from that, to the best of our knowledge, there is no similar approach in the literature that considers the use of SVM to perform path loss prediction. Thus, in this work we propose a method for predicting path loss in an urban outdoor environment using support vector regression (SVR). Simulation results indicate that, depending on the employed kernel and its parameters, the performance obtained using support vector regression is similar and with lower computational complexity to that obtained by a MLP neural network.

The remainder of this article is structured as follows. In Section II, two empirical propagation models are presented: Okumura-Hata model and Ericsson 9999 model. Concepts about support vector regression and the measurement setup are described in Section III. Section IV presents the model tuning of the SVR techniques and numerical results. At last, conclusions are drawn in Section V.

II. EMPIRICAL PROPAGATION MODELS

Reliable and accurate propagation models are crucial to the prediction of radio channel characteristics for where the wireless network system is to be deployed. In general, propagation models can be categorised into two types: deterministic and empirical.

Deterministic propagation models consider the physical paths along which the transmitted waves propagate are usually based on ray optical techniques [9]. These models describe wave propagation by different rays that travel from the transmitting to the receiving antenna and are subjected to reflection, scattering and diffraction at walls and edges of buildings and similar objects. Deterministic models offer excellent accuracy and are able to provide additional parameters such as small-scale fading, delay spread, etc. The main disadvantage of the deterministic models is their large computation time.

On the other hand, empirical propagation models are those based only on observations and measurements. In spite of these

ISBN: 978-1-61208-360-5

models be able to predict rain-fade and multipath [10], they are mainly used to estimate path loss, an important task during the initial deployment of wireless networks and cell planning. Empirical models can be split into two subcategories namely time dispersive and non-time dispersive. Time dispersive models are designed to provide information relating to the time dispersive characteristics of the channel. An example of this type are the Stanford University Interim (SUI) channels models developed under the IEEE 802.16 working group [11]. On the other hand, a non-time dispersive model predicts mean path loss as a function of various parameters as distance, antenna heights, latitude, longitude, etc. Examples of non-time dispersive empirical models are Hata [12] and Ericsson 9999 [13] models. In this work, we consider the Okumura-Hata and Ericsson 9999 models for path loss prediction.

A. Okumura-Hata Model

The Okumura model is one of the most used models for signal prediction in the urban areas. It applies to frequencies in the range of 150 MHz to 1920 MHz and distances from 1 to 100 km [14]. The Okumura model is entirely based on measured data without any analytical explanation. Nevertheless, it is very practical and has become a standard for planning land mobile radio systems in Japan. The Okumura model is a very good model in urban and suburban areas, but not so good in rural areas due to its slow response to rapid changes in the terrain.

The Hata model is an empirical formulation of the path loss data provided by Okumura's model and it is valid from 150 MHz to 1500 MHz. Hata presented the propagation loss in urban area as a standard expression and provided correction factors for applications in other environments. The Okumura-Hata model is the combination of both above models.

The expression for the average path loss in urban areas is given by [12]

$$\begin{aligned} L(\mathrm{dB}) &= 69.55 + 26.26\log f - 13.82\log h_t \\ &- a(h_r) + (44.9 - 6.55\log h_t)\log d \qquad (1) \end{aligned}$$

where f is the frequency (in MHz) from 150 MHz to 1500 MHz, h_t is the effective height of the base station antenna (transmitter) in meters, varying from 30 m to 200 m, h_r is the effective height of the mobile station antenna (receiver) in meters, varying from 1 m to 10 m, d is the distance between transmitter and receiver (in km), and $a(h_r)$ is the correction factor to the effective height of the receiver antenna, which is function of the size of the coverage area. For large cities and $f \geq 300$ MHz, the factor $a(h_r)$ is given by

$$a(h_r) = 3.2(\log 11.75h_r)^2 - 4.95 \text{ dB} . \qquad (2)$$

For suburban and rural areas, the path loss is obtained by other expressions that can be found in [15].

Predictions of the Hata model are quite similar to the Okumura model, whereas d does not exceed 1 km. The Hata model applies to macrocells mobile systems, but not to personal communications service (PCS) systems that have cells in order of 1 km radius.

B. Ericsson 9999 Model

The Ericsson 9999 model is an extension of Hata model, where we can adjust the parameters according to the given scenario [13]. In this model, the path loss is described as

$$\begin{aligned} L(\mathrm{dB}) &= a_0 + a_1\log d + a_2\log h_t + a_3\log h_t\log d \\ &- 3.2(\log 11.75h_r)^2 + g(f) \qquad (3) \end{aligned}$$

where $g(f)$ is given by [13]

$$g(f) = 44.49\log f - 4.78(\log f)^2 . \qquad (4)$$

For urban environments, the default values of the parameters a_0, a_1, a_2 and a_3 are, respectively, 36.2, 30.2, 12.0 and 0.1. For suburban and rural environments, the parameters a_0, a_1, a_2 and a_3 assume another values that can be found in [13].

III. Proposed Method

SVM is a popular ML technique that make use of the optimization of a function in its training stage. Lately, SVM have received increasing attention from ML community, since it presents some advantages when compared with other ML techniques, such as the absence of local minima, the sparseness of the solution and the capacity control obtained by optimising the margin [8]. Initially developed for solving classification problems, SVM techniques can be successfully applied in regression, i.e., for function approximation problems.

A. Support Vector Regression

In the regression problems, we estimate the functional dependence of the output variable on an n-dimensional input variable. In other words, we deal with real valued functions and we model an $\mathbb{R}^n$ to $\mathbb{R}$ mapping.

The general regression learning problem is set as follows. The ML algorithm is given a set of training data from which it tries to learn the input-output relationship. So, consider a training data set $D = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \mathbb{R}, i = 1, 2, \ldots, \ell\}$ with ℓ pairs $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_\ell, y_\ell)$, where the inputs are n-dimensional vectors $\mathbf{x}_i \in \mathbb{R}^n$, the outputs $y_i \in \mathbb{R}$ are continuous values and ℓ is the number of samples in the training data set.

Starting from the linear regression problem, assume that $h(\mathbf{x}_i, \mathbf{w})$ is a linear regression hyperplane given by

$$h(\mathbf{x}_i, \mathbf{w}) = \langle \mathbf{w}^T, \mathbf{x}_i \rangle + b \ , \qquad (5)$$

where $\mathbf{w} \in \mathbb{R}^n$ is the normal vector to this hyperplane, the scalar $b \in \mathbb{R}$ is called a bias, $\langle\cdot,\cdot\rangle$ is the inner product operator and $(\cdot)^T$ is the transpose operator. In the case of SVR, we measure the error of approximation instead of the margin used in classification. With this in mind, we use a function named Vapnik's linear loss function with ε-insensitivity zone defined as [16]

$$E(e_i) = |e_i|_\varepsilon = \begin{cases} 0 & , \text{ if } |e_i| \leq \varepsilon \\ |e_i| - \varepsilon, & \text{otherwise} \end{cases} \ , \qquad (6)$$

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-360-5

where $e_i = y_i - h(\mathbf{x}_i, \mathbf{w})$. The Vapnik's linear loss function $E(e_i)$ is illustrated in Fig. 1(a), where the ε-insensitivity zone is highlighted. Thus, the loss is equal to zero if the difference between the predicted $h(\mathbf{x}_i, \mathbf{w})$ and the measured value y_i is less than ε.

The solution of the linear regression learning problem concerns to find the linear function that approximates the training pairs $(\mathbf{x}_i, y_i)$ with an accuracy ε. In other words, we need to find a vector $\mathbf{w}$ that minimizes the error, which implies to solve the optimization problem given by [17]

$$\min_{\mathbf{w},b} \frac{1}{2} \|\mathbf{w}\|^2 \tag{7}$$

restricted to $|e_i| \leq \varepsilon$.

To obtain sparse solutions and penalize the large residuals, a penalty term is included in (7), so that

$$\min_{\mathbf{w},b} \left[\frac{1}{2} \|\mathbf{w}\|^2 + C \left(\sum_{i=1}^{\ell} E(e_i) \right) \right] \tag{8}$$

where C is a cost parameter. The function $E(e_i)$ defines an ε-tube as exhibited in Fig. 1(b), where ε is the radius of the tube. The restriction $|e_i| \leq \varepsilon$, i.e., $y_i + \varepsilon \geq h(\mathbf{x}_i, \mathbf{w}) \geq y_i - \varepsilon$ is the condition for a predict point to be within in the ε-tube.

The optimization problem represented by (8) can be relaxed by introducing slack variables, denoted by ξ and $\hat{\xi}$, which allows to deal with points outside the ε-tube. The points above the ε-tube have $\xi > 0$ and $\hat{\xi} = 0$, while the points below the ε-tube have $\xi = 0$ and $\hat{\xi} > 0$. At last, the points inside of ε-tube have $\xi = \hat{\xi} = 0$.

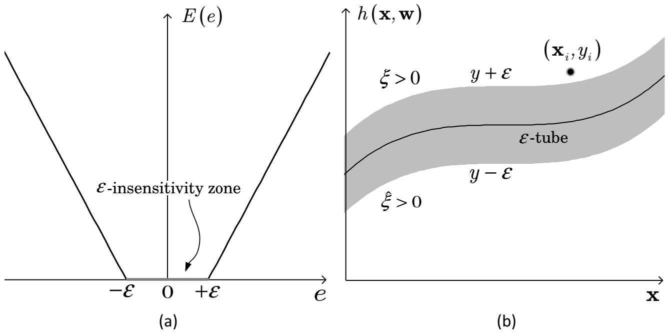


Fig. 1. (a) Vapnik's linear loss function with ε-insensitivity zone versus e. (b) ε-tube defined from the function $E(e)$.

Given the slack variables ξ and $\hat{\xi}$, we can rewrite the optimization problem as

$$\min_{\mathbf{w},b} \left[\frac{1}{2} \|\mathbf{w}\|^2 + C \left(\sum_{i=1}^{\ell} (\xi_i + \hat{\xi}_i) \right) \right] \tag{9}$$

under the restrictions

$$\left\{ \begin{array}{c} |e_i| = \varepsilon + \xi \\ |e_i| = \varepsilon + \hat{\xi} \\ \xi, \hat{\xi} \geq 0 \end{array} \right. ,$$

which can be solved using Lagrange multipliers, as can be seen in [17]. After calculating the Lagrange multiplier vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$, the best regression hyperplane obtained is given by

$$h(\mathbf{x}_i, \mathbf{w}) = \sum_{i=1}^{\ell} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \left\langle \mathbf{x}_i^T, \mathbf{x}_i \right\rangle + b \ . \tag{10}$$

In the case of nonlinear regression, the basic idea is to map the input vectors $\mathbf{x}_i \in \mathbb{R}^n$ into vectors $\Phi(\mathbf{x}_i)$ of a higher dimensional feature space $\Im$, where Φ represents the mapping. After this transformation, a nonlinear problem in $\mathbb{R}^n$ becomes a linear problem in the feature space $\Im$. So, the optimization problem is reformulated as the maximization of dual Lagrangian with Hessian matrix [8] and the solution is given by

$$h(\mathbf{x}_i, \mathbf{w}) = \sum_{i=1}^{\ell} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \left\langle \Phi^T(\mathbf{x}_i), \Phi(\mathbf{x}_i) \right\rangle + b \ . \tag{11}$$

in which the summation is not performed over all training data, but rather over those that have non-zero Lagrange multipliers, which are called *support vectors*.

Note that the optimization problem for nonlinear regression, represented by (11), involves the calculation of inner products between vectors of the feature space $\Im$. Since $\Im$ can be very higher dimensional, the calculation of Φ can become infeasible. Therefore, the solution is to resort to the *kernel trick*, i.e., the use of kernels to perform nonlinear regressions without mapping all input vectors $\mathbf{x}_i$ to the feature space $\Im$ [18].

A kernel is a function that applies to two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ in the input space X and returns the inner product of these vectors in the feature space $\Im$ [19], i.e.,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \ . \tag{12}$$

To ensure the convexity of the optimization problem given by (11) and that the kernel represents mappings in which it is possible the calculation of the inner products $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$, kernel functions satisfying the conditions of Mercer are exploited [16]. The more common practice kernels for regression problems are the polynomial kernel and the radial basis functions (RBF) ones. In this paper, we consider the use of the polynomial kernel and two types of RBF kernels: Laplacian and Gaussian. The expressions related to each kernel are given in Table I.

TABLE I
TYPES OF KERNELS CONSIDERED FOR THE PROPOSED METHOD.

Kernel	Expression	Parameters
Polynomial	$K(\mathbf{x}_i, \mathbf{x}_j) = \left(\beta \left\langle \mathbf{x}_i^T, \mathbf{x}_j \right\rangle + c\right)^z$	β, c, z
Gaussian	$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$	σ
Laplacian	$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{\sigma}\right)$	σ

The polynomial kernel is a function that represents the similarity of training samples in the feature space over polynomials of the original variables and combinations of these. The adjustable parameters are the scale term β, the constant term (off-set) c and the polynomial degree z. Laplacian and Gaussian kernels are examples of RBF kernels. The adjustable parameter σ is very important to the performance of these kernels and should be fit to the problem at hand.

B. Measurement Setup

The work presented in this paper considers mobile radio wave propagation measurements at a carrier frequency of 853.71 MHz. The measurements of the downlink signal strength level were made in an urban environment in the city of Fortaleza-CE, Brazil. Fig. 2 illustrates the urban area of the city where the measurements were taken. The colors used along the indicated streets represent each received signal strength indicator (RSSI) in dBm. In total, 1933 measurements were performed using Agilent E6474A tool as a pilot scanner. The location of the base station (BTS) is also indicated in Fig. 2.

During the drive test, various field data of each measured point were collected to compose the feature vector of the SVR process. Such field data were antenna-separation distance, terrain elevation, horizontal angle, vertical angle, latitude, longitude, horizontal and vertical attenuation of the antenna. At last, the theoretical path loss of the Okumura-Hata model was also used as an input of the SVR training algorithm. The terrain elevation was collected using Google Elevation API by a Java client made exclusively for it. The base station was located in a rooftop 90 meters high with a sectored antenna, which had a half-power beam width of 63 °. Also, the effective radiated power (ERP) of the base station was set to 48 dBm.

IV. Training and Evaluation

Many ML algorithms, such SVR and ANN, have important parameters that cannot be set directly from the data. The process of choosing these parameters to obtain the best performance of the model is known as tuning and is described below.

A. Model Tuning

Cross-validation is a model validation technique for evaluating how the results of a statistical analysis will generalize to an independent data set. A common type of cross-validation is the k-fold cross-validation, generally used to evaluate the model accuracy [20]. It is a re-sampling technique where the samples are randomly split into k sets of approximately equal size. These subsets are named folds and they are divided in two groups: the test set with only one fold and the training set with $(k-1)$ folds. Initially, the first fold is established as test set and the model is fit using the others $(k-1)$ folds. The held out sample in the first fold is predicted by the ML algorithm and is utilized to estimate the performance. After that, the first fold is given back to the training set. This procedure is repeated with the second fold held out, and so on. In this paper, we consider $k = 10$ and use the average root mean square error (RMSE) $\bar{\mu}$ defined as

$$\bar{\mu} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\mu_j} \tag{13}$$

to evaluate the model precision. In (13), μ_j is the RMSE calculated for j-th test set ($j = 1, 2, \ldots, k$), given by

$$\mu_j = \sqrt{\frac{1}{\ell_j}\sum_{i=1}^{\ell_j}\left(y_i - h\left(\mathbf{x}_i, \mathbf{w}\right)\right)^2} \tag{14}$$

where ℓ_j is the number of samples in the j-th test set.

The definition of σ, common to RBF kernels, was made using the analytical approach presented in [21], where it is shown that the optimal values of σ are in the range of the 10^{th} and the 90^{th} percentile of $\|\mathbf{x}_i - \mathbf{x}_j\|^2$. In addition to that, it is suggested in [18] that the midpoint of these two

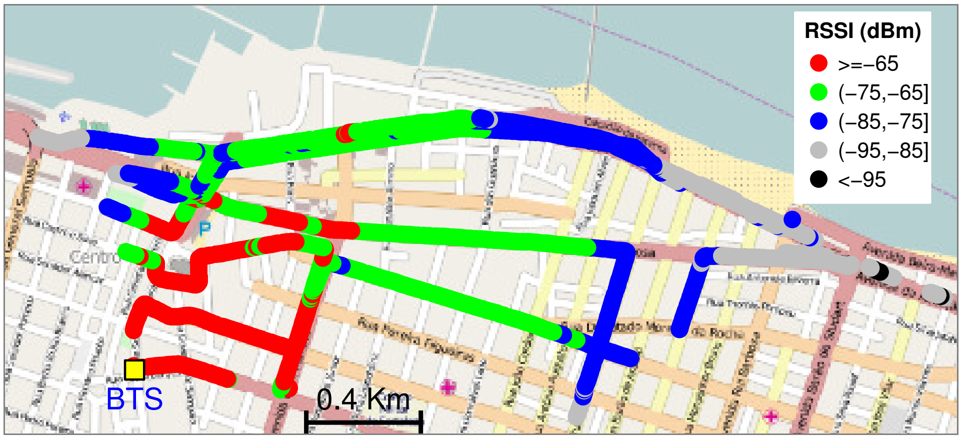


Fig. 2. Drive test with measurements in an urban area in the city of Fortaleza. The colors indicate the radio signal strength indicator (RSSI) in dBm and the yellow square represents the location of the base station (BTS).

percentiles should be used. Thus, this kernel parameter was estimated to be $\sigma = 0.244$.

The cost C, common parameter to all kernels, is fundamental for adjusting the complexity of the model. When the cost is large, the model is more flexible, but it becomes more likely to over-fit. With a small cost, the flexibility of the model decreases, but the over-fit is less likely. However, a small cost can lead to poor predictions due to under-fit [20]. In the tuning process, 18 values were tested for C, from 2^{-2} to 2^{15}, being each value a power of 2.

In the tuning process considered in this work, it was tested $\varepsilon = 0.1$ and $\varepsilon = 0.05$ in combination with the range of C specified previously. The best fit was found for $\varepsilon = 0.05$ to the Laplace kernel, and $\varepsilon = 0.1$ to the Polynomial and the Gaussian ones.

B. Numerical Results

All models considered in this paper are implemented using the R language. The performance of the SVR algorithms is evaluated via computer simulations for the three kernels mentioned in Subsection III-A. For their implementation, the kernlab package is employed [18].

According to what has been explained about SVR for nonlinear learning problems in Subsection III-A, Fig. 3 shows the support vectors and the respective regression line when the Laplace kernel is adopted. Note that, for the sample set in evidence, the number of support vectors is inferior to the number of measurements, but it is sufficient to obtain the regression line.

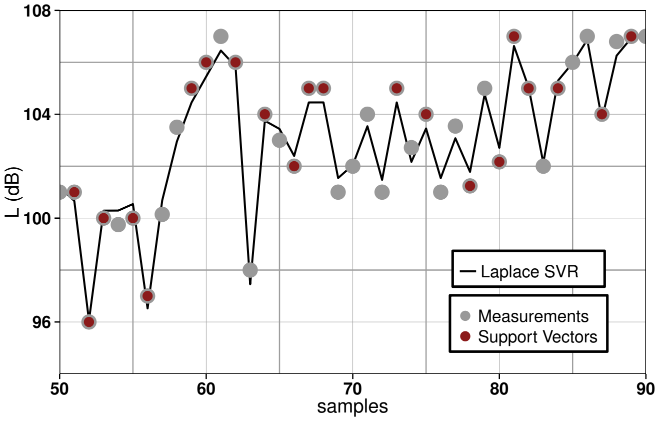


Fig. 3. Support vectors used to obtain the regression line for path loss in dB when Laplacian kernel is adopted.

Fig. 4 shows a comparison of measurements corresponding to 10% of the sample set obtained from the drive test and the predictions using SVR algorithms for polynomial, Gaussian and Laplacian kernels. One can see that the Laplacian kernel is the best option among the three kernels.

For comparison of the SVR algorithms with other ML techniques, it is implemented a MLP-ANN with a weight decay w_d, having a input layer with nine neurons, a hidden layer with M neurons and a output layer with one neuron. The backpropagation algorithm is used to train the MLP-ANN. In the tuning process, some MLP-ANN configurations were investigated for $M = 9, 12, 15, 18, 21, 24, 27$ and $w_d = 0.01, 0.05, 0.1$. The best fit found for the MLP-ANN was $M = 27$ and $w_d = 0.01$.

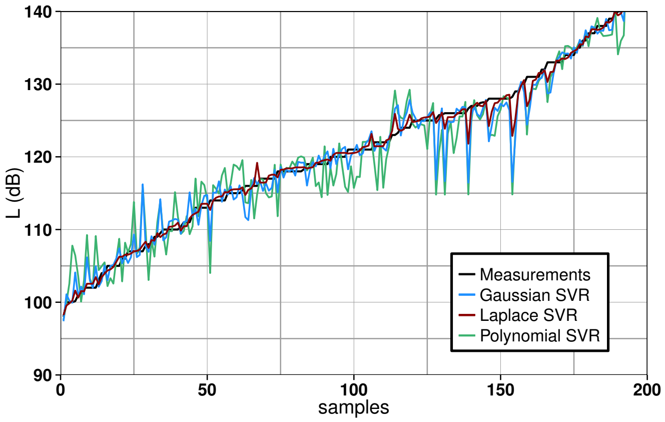


Fig. 4. Comparison of path loss predictions using SVR algorithms and measurements obtained from drive test for 10% of the sample set.

Table II provides a statistical analysis of the SVR algorithms, the MLP-ANN and the two empirical propagation models mentioned in Section II, where the average RMSE $\bar{\mu}$ and the standard deviation of the RMSE, denoted by σ_μ are presented. The best configuration parameters of the SVR algorithms (C and ε) and the parameters of each kernel are also shown in Table II. We can see that the Laplacian SVR presents an average RMSE $\bar{\mu} = 1.76$ dB, while the polynomial and the Gaussian SVRs presents an average RMSE of 3.47 dB and 4.55 dB, respectively. Both empirical propagation models have inferior performance when compared to all considered ML techniques.

The MLP-ANN performance, with $\bar{\mu} = 1.89$ dB, can be considered similar to the Laplacian SVR performance. In spite of analogous performance, MLP-ANN and Laplacian SVR have significant differences in their implementation, which are discussed below.

At first, as the MLP-ANN has local minimal, it is necessary to initialize the weight matrix with different values in the attempt to test more points (in this work, we initialize the MLP-ANN three times), whereas such problem do not exist in SVR algorithms. In case of SVR, a convex optimization problem is solved resulting in a global minimum. Therefore, when using SVR there is no problem with initializations and checking for convergence [8].

Secondly, as there is no feature extraction in the MLP-ANN, sometimes it is necessary to use another ML algorithm to do this task [18]. In the SVR algorithm, the data can be applied directly without the need for feature extraction, because the SVR algorithm already do this function [22]. Furthermore, when the number of features increases, the MLP-ANN complexity demands more computational cost,

TABLE II
STATISTICAL ANALYSIS OF THE SVR ALGORITHMS WITH THEIR BEST PARAMETERS C AND ε, MLP-ANN AND THE EMPIRICAL MODELS CONSIDERED IN THIS WORK.

ML Algorithm/Model	$\bar{\mu}$ (dB)	σ_μ (dB)	C	ε	Kernel parameters
Polynomial SVR	3.47	0.54	1024	0.1	$\beta = 0.1, c = 1, z = 3$
Gaussian SVR	4.55	0.15	8192	0.1	$\sigma = 0.244$
Laplacian SVR	1.76	0.12	1024	0.05	$\sigma = 0.244$
MLP-ANN	1.89	0.17	-	-	-
Okumura-Hata	7.13	5.08	-	-	-
Ericsson 9999	21.59	6.22	-	-	-

whereas in the SVR algorithm, once that a valid kernel has been selected, one can practically work in spaces of any dimension without any significant additional computational cost [22].

Finally, the MLP-ANN training time is normally longer than SVR one. We can mention two reasons for that: first, the MLP-ANN usually needs to be initialized more than one time; second, in the SVR algorithm, the training is executed considering only the support vectors, while in the MLP-ANN the training is performed on the entire data set.

V. CONCLUSIONS

In this study, a method to predict path loss in an urban outdoor environment using SVR was proposed. To do that, mobile radio wave propagation measurements at a carrier frequency of 853.71 MHz obtained in an urban environment in the city of Fortaleza-CE, Brazil were considered. Various field data of each measured point like antenna-separation distance, terrain elevation, the theoretical path loss of the Okumura-Hata model among others were collected as input of the SVR process. Polynomial, Gaussian and Laplacian kernels were adopted for SVR algorithms. For comparison, we considered two empirical propagation models (Okumura-Hata and Ericsson 9999) and a MLP-ANN optimized for our prediction problem. In case of SVR, it was verified that the Laplacian kernel was the best option among the investigated kernels. In addition, the SVR algorithm using Laplacian kernel and the MLP-ANN had similar performance, being the former an alternative of lower computational complexity. The authors conjecture that the lower computational complexity of the SVR technique is due to the use of support vectors and the kernel trick which reduce the training time, naturally perform the feature extraction and increase the capacity of working with higher dimensional spaces.

ACKNOWLEDGMENTS

The authors would like to thank Evandro Uchoa and Gustavo Raulino for their assistance in the acquisition of field data.

REFERENCES

[1] R. Want, "When cell phones become computers," *IEEE Pervasive Computing*, vol. 8, n. 2, 2009, pp. 2-5.

[2] T. S. Rappaport. *Wireless Communications: Principles and Practice*. Prentice-Hall PTR, 2 ed., 2009.

[3] T. Mitchell. *Machine Learning*. McGraw-Hill, 1 ed., 1997.

[4] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, NJ, USA, 2 ed., 1999.

[5] E. Östlin, H.-J. Zepernick and H. Suzuki, "Macrocell path-loss prediction using artificial neural networks," *IEEE Vehic. Tech.*, vol. 59, n. 6, 2010, pp. 2735-2747.

[6] P. S. Sotiroudis et. al., "A neural network approach to the prediction of the propagation path-loss for mobile communications systems in urban environments," in *Proc. of the Progress in Electromag. Research Symp. (PIERS)*, Prague, CZ, 2007.

[7] A. Neskovic et. al., "Indoor electric field level prediction model based on artificial neural networks," *IEEE Commun. Lett.*, vol. 4, n. 6, 2000, pp. 190-192.

[8] V. Kecman, *Support Vector Machines: An Introduction*. Springer, New York, 2005.

[9] G. E. Athanasiadou et. al., "A microcellular ray-tracing propagation model and evaluation of its narrowband and wideband predictions," *IEEE J. Sel. Areas Commun.*, vol. 18, n. 3, 2000, pp. 322-335.

[10] R. K. Crane, "Prediction of attenuation by rain," *IEEE Trans. Commun.*, vol. COM-28, 1980, pp. 1727-1732.

[11] V. Erceg et. al., "Channel models for fixed wireless applications," *Tech. Rep.*, IEEE 802.16 Broadband Wireless Access Working Group, Jan 2001.

[12] M. Hata, "Empirical formula for propagation loss in land mobile radio services," *IEEE Trans. Veh. Tech.*, vol. 29, n. 3, 1981, pp. 317-325.

[13] S. S. Kale and A. N. Jadhav, "Performance analysis of empirical propagation models for WiMAX in urban environment," *IOSR Journal of Electronics and Communication Engineering*, 2013, pp 24-28.

[14] Y. Okumura et. al., "Field strength and its variability in VHF and UHF land-mobile radio-services," *Rev. Elec. Comm. Lab.*, vol. 16, Sep-Oct 1968.

[15] T. K. Sarkar et. al., "A survey of various propagation models for mobile communication," *IEEE Ant. and Propag. Mag.*, vol. 45, n. 3, 2003, pp. 51-82.

[16] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[18] A. J. Smola, K. Hornik and A. Karatzoglou. "An S4 Package for Kernel Methods in R," *Journal of Statistical Software*, vol.11, n.9, 2006, pp 1-20.

[19] R. Herbrich. *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, 2001.

[20] M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer, New York, 2013.

[21] B. Caputo, K. Sim, F. Furesjo and A. Smola. Appearance-based Object Recognition using SVMs: Which Kernel Should I Use?, in *Proc. of NIPS Workshop on Statist. Methods for Comput. Exp. in Visual Process. and Comp. Vision*, Whistler, 2002.

[22] K. Hornik, D. Meyer and K. Hornik. "Support vector machines in R," *Journal of Statistical Software*, vol. 15, n. 9, 2006, pp 1-28.

Applicable Cost Modeling of LTE-Advanced and IEEE 802.11ac based Heterogeneous Wireless Access Networks

Vladimir Nikolikj
Vip operator -
Member of Telekom Austria Group
Skopje, Macedonia
E-mail: v.nikolikj@vipoperator.mk

Toni Janevski
Ss. Cyril and Methodius University
Faculty of Electrical Engineering and Information
Technologies, Skopje, Macedonia
E-mail: tonij@feit.ukim.edu.mk

Abstract—**In this paper, we propose applicable and comparative cost-capacity analysis of the heterogeneous wireless networks in order to determine the most cost effective radio network deployment strategies as a function of an extreme demand levels of even more than 100 GB per user and month. We perform the modeling by considering of the unit cost drivers relevant for the various base station classes which provide different coverage and high capacity performance, coming with the Long Term Evolution Release 10 (LTE-Advanced) radio access technology or IEEE 802.11ac Wi-Fi standard. Considering different amounts of available bandwidth in the 800 MHz and 2.6 GHz bands, the key finding is that the small cell solutions like femto cells and Wi-Fi are more cost efficient when new macro base station sites need to be deployed or when very high demand levels need to be satisfied. In all other evaluated cases, the importance of the spectrum size comes to the highest level together with the introduction of the LTE-Advanced carrier aggregation functionality. Also, we evaluate the economic gains of a joint deployment of femto/Wi-Fi sites from one side and macrocells from other side. We determine that instead of investing in additional spectrum or deploying denser macro network, mobile operators could compensate the indoor wall penetration losses by deploying different number of femto sites per floor or user per femto site, for still satisfactory level of QoS.**

Keywords-Wireless Heterogeneous Networks; Cost modeling; LTE-Advanced; IEEE 802.11ac.

I. INTRODUCTION

The rapid increase of mobile broadband services has resulted in a marvel of decoupling the traffic load from operator revenues. Flat service subscriptions nowadays, even further increases the challenge of the Mobile Network Operators (MNOs) to monetize on the data traffic. Hence, it is from the highest importance of the MNOs to deploy more cost effective networks that will respond to the increasing user demand. The forthcoming wireless network architectures become more heterogeneous, with hierarchically ranged Base stations (BS) sites/cells, as follows: macro (MaBS) to cover wider areas, and micro (MiBS), pico (PBS) and femto (FBS) complemented with particular wireless local area network (WLAN/Wi-Fi) to cover smaller areas. A number of papers have been published on modeling the cost-effectiveness by comparing the MaBS cell deployment with the small-cell deployment and suggesting utilization of joint or heterogeneous and even cooperative networks. Analysis of MaBS, MiBS and PBS HSPA cells capacity-cost comparisons including IEEE 802.11a, are provided in [1][2][3]. Cost comparisons of LTE with HSPA deployed MaBS networks and FBS solutions are extensively covered in [3][4]. Additionally, the evaluation of the economic gain provided by various deployments of FBS and MaBS for LTE mobile broadband services is outlined in [5].

In this article, we originally introduce the comparative cost modeling of MaBS, MiBS, PBS and FBS utilizing LTE-Advanced (LTE-A) RAT [6][7], alongside with Wi-Fi standard IEEE 802.11ac [8]. Considering the “up to date” initial and running cost drivers, together with the coverage and capacity specific parameters, we deliver results helping more easily to assess the most cost efficient manner to deploy the heavily-loaded, wireless heterogeneous networks. The special focus is put on the comparison between MaBS and various small cell deployments. As according to Analysis Mason [9], more than 80% of the mobile traffic is generated in indoors, we create long-term investment case study related to indoor office users. In order to determine more realistic cost-capacity performance modeling, besides already discussed wall attenuation and indoor coverage strategies in [3][4], additionally, we consider the performance of the carrier aggregation functionality of LTE-A RAT. For all deployment scenarios, we analyze the deployment of new sites and reusing the existing sites.

Still having in mind that each cellular network in reality consists of a mix of BSs, we conduct the joint heterogeneous network cost-capacity analysis as well. The assessment of the economic gains of joint deployments is done for the period of 10 years by using the discounted cost model in order to take into account the “time value of investment and running costs”.

The paper is organized as follows. Sections II and III describe the analysis approach through elaboration of RAN specific coverage, capacity and unit cost estimates for various BS classes. In the next section, we perform investment modeling of various wireless network deployment strategies through the case study. Based on the results from Section IV, in Section V, we discuss the findings and analyze the most and less cost-effective scenarios separately deployed. In Section VI, we demonstrate the combined cost-capacity modeling of different wireless heterogeneous network solutions to satisfy high demand levels. A conclusion is drawn in Section VII.

ISBN: 978-1-61208-360-5

II. Radio Access Network Coverage and Capacity Modeling

Consisted of numerous BS sites and Radio Network Controllers (RNCs), the Radio Access Network (RAN) of the MNO is deployed to provide services within the entire system coverage area denoted as A_{syst}. According to Johansson et al. [10], a BS of class *i* is characterised by a maximum average throughput or capacity T_{maxi} and cell range r_i related to coverage. Based on the purpose of use, a BS of class *i* could be equipped with radio equipment supporting up to three sectors and up to three different frequency carriers. The number of cells $N_{cel,}$ within the particular BS site *i*, is obtained as multiple of the number of supported sectors and frequency carriers. We model the coverage A_{cell} of a particular cell area of BS site *i* as follows:

$$A_{cell} = \pi \cdot {r_i}^2 \qquad (1)$$

The maximum path loss allows the maximum cell range to be estimated with a suitable propagation model, such as Okumura-Hata [12]. Based on [11], the calculation shows that urban cell range varies from 0.6 km at 2.6GHz to 1.4 km at 900 MHz. Since in this paper we focus on the urban dense area, according to Markendahl and Mäkitalo [3] and Markendahl [4], we consider 0.57 km range for MaBS. Based on the elaborations in [1][2], we estimate 0.27 for MiBS and 0.1 km range for PBS. FBS cell range in [3] is assumed at 0.050 km and in [13] in range of 0.01 – 0.030 km. According to Mölleryd et al. [14], we model the aggregated capacity of the system, T_{syst}, as follows:

$$T_{syst} = W \cdot N_{site} \cdot N_{cell} \cdot S_{eff} \qquad (2)$$

where *W* is allocated bandwidth in MHz, N_{site} is the total number of BS sites within the system coverage area A_{syst} and S_{eff} is the cell average cell spectral efficiency in bps/Hz/cell. Based on [6] [7] the average spectral efficiency for LTE-A varies from 6.6, 4.2 and 3.8 bit/s/Hz/cell for the indoor, microcellular and base coverage urban environments, respectively (environments are determined in line with [15]). With regard to the FBS deployment, interference problems to non-FBS cell occur with the creation of the so called “Closed User Group” deployment FBS model [16]. As proposed by [17], in adjacent-channel deployments (the FBS is deployed on a dedicated carrier), the coverage holes are considerably easier to minimize and control than when the FBS is deployed on the same carrier as the macro layer (co-channel deployment - sharing the channel with the MaBS network). Hence, in this article we consider FBS deployment in a different frequency band than MaBS. Currently, the LTE FBS are developed with 5, 10 and 15 MHz bandwidth (achieving up to 37, 75 and 112 Mbps in downlink, respectively) and available from 8 to 16 users simultaneously [18]. Choi in [13], indicates that 4G FBS will utilize the bandwidth of 20 MHz per carrier. We use the indoor average spectral efficiency of 6.6 bps/Hz and 20 MHz of spectrum for FBS with 50m coverage range. According to Xiao [19], it is very difficult to exceed 50-60% of the nominal bit rate of the underlying physical layer of Wi-Fi. Frame aggregations techniques are used to improve the Medium Access Control (MAC) layer efficiency [20].

TABLE I. Radio Access Network - Coverage and Capacity Parameters.

BS Parameter/ LTE-A and IEEE 802.11ac	MaBS	MiBs	PBS	FBS	Wi-Fi
Range (km)	0.57	0.25	0.10	0.05	0.03
Coverage (km²)	1.02	0.19	0.03	0.008	0.003
Sectors	3	1	1	1	1
Carriers	1 – 3	1 – 2	1	1	1
Cells	3 – 9	1 – 2	1	1	1
Bandwidth (MHz)	20	20	20	20	80
Av. Cell SE (bps/Hz)	3.8	3.8	6.6	6.6	16.25
Av. Cell Capac.(Mbps)	76	76	132	132	1300
Av. Site Capac. (Mbps)	228	76	132	132	1300

According to Cisco [21], we consider the first-wave IEEE 802.11ac products operating in the 5 GHz band with 80 MHz and delivering up to 1300 Mbps (high end) at the physical layer up to 30 m coverage range. Based on the all above coverage and capacity estimates, we summarize in Table I, RAN coverage and capacity parameters related to different RAT as used in this paper.

III. Heterogeneous Radio Access Network Cost Modeling

We base our cost structure modelling to the methodology developed in [1] [5] by limiting to the capital investment to acquire and deploy the RAN (CAPEX), and the costs to operate the RAN (OPEX). We consider the BS equipment, BS site installation & buildout, backhaul transmission equipment and Radio Network Controller equipment as BS related CAPEX items and electric power, operation & maintenance, site lease and backhaul transmission lease as BS related OPEX items. Also, we evaluate the CAPEX and OPEX of system spectrum. Regardless that CAPEX consists of one-time expenditures, usually for practical reasons these expenditures are spread over several years, i.e., annualized. Still, according to METIS [22] an even more accurate model could be obtained by using present values instead of annualizing the CAPEX. In order to calculate the cost per item of type *i* in present value, according to Johansson et al. [2], we use the standard economical method for cumulated discounted cash flows yield by summing up the total discounted annual expenditures for the whole network life cycle (K years) as follows:

$$\mathcal{E}_i = \sum_{k=0}^{K-1} \frac{\alpha_{k,i}}{(1+\beta)^k} \qquad (3)$$

where $\alpha_{k,i}$ is the sum of expenditures, in terms of CAPEX and OPEX, occurred within year k of an item of type *i* and β is the discount rate. In all analyzed scenarios in this paper, we assume network life cycle of K = 10 years and that all BSs are installed during the first year of deployment. Additionally, according to Frias and Pérez [5], we use the discounted rate equalized to the cost of capital (a WACC - weighted average cost of capital) of β = 12 %. Consequently, the total discounted cost, C_{TOT}, of a wireless heterogeneous access network comprising of macro sites and small sites normalized per unit of area, can be approximated as follows:

$$C_{TOT} = \mathcal{E}_M \cdot N_M + \mathcal{E}_S \cdot N_S + \frac{\mathcal{E}_{SPECTRUM}}{A_{syst}} \quad [cost/area] \qquad (4)$$

where $\mathcal{E}_M$ *is the* total discounted cost of MaBS, $\mathcal{E}_S$ *the* total discounted cost of small BS (or Wi-FI BS), $\mathcal{E}_{SPECTRUM}$ is the total discounted cost for spectrum licenses, A_{syst} is the coverage area of entire operator's network and N_M and N_S is the average number of MaBSs and small BSs, respectively (in this paper, we will not consider the costs that are not related to the technical solution, such as customer care and marketing, as well as the average customer retention or subsidy costs). The cost estimates related to different RATs and BS classes will be derived in the next subsections.

A. Base Stations Unit Cost Estimates

The cost per BS is significantly different based on the BS type considered in this paper. For example, BSs providing bigger coverage imply a higher cost for equipment, site leases and installation whereas a small cells BS sites costs much less in those aspects. Nevertheless, according to Johansson et al. [2], fixed costs not directly related to the capacity of the BS are divided between many users in a MaBS so the cost per user may still be lower in many scenarios.

According to Markendahl and Mäkitalo [3], the estimates for year 2010 show that cost for deploying a new MaBS site in the urban area is 110 k€ including transmission and that the cost for radio equipment supporting three sectors and 5–20 MHz to 10 k€, yielding to total CAPEX of 120 k€. According to Johansson [1], CAPEX for 2-carrier and single carrier MaBS deployment is 20% and 40% lower than 3-carrier MaBS, respectively. Notice though that the costs for an LTE-A cellular network are hypothetical since the system is now being released to the market. Out of Johansson [1], we consider the price of a MiBS and PBS station equals 50% and 15%, respectively, of a single-carrier MaBS equipment, with a note that PBS needs 2 k€ for transmission, and MiBS and PBS requires 10 k€ and 2 k€ for the site deployment, respectively. According to Markendahl [4], on average the deployment of one FBS is around 1 k€.

Even prior the full standardization, some manufactures start to offer IEEE 802.11ac products. WLAN Access points (AP) for consumers are currently available at prices of around €160 [24]. Nevertheless, for the enterprise solutions there should be used WLAN carrier grade access [25] [26]. Johansson in [1], outlines that the carrier grade AP is 10 time more expensive than WLAN AP for consumers, and that cost for router and access getaway is 20 k€. Consequently, we assume that carrier grade access point supporting IEEE 802.11ac will cost around 1.5 k€, and additional 1k€ should be added per AP, assuming that the control equipment is divided between 20 APs.

Regarding the OPEX, Markendahl and Mäkitalo in [3] assume 30 k€ annual cost for the new MaBS site and Johansson in [1] considers 13.4 k€ for the single carrier MaBS by outlining an appropriate ratios of 1.15, 1.29, 0.67, 0.21 and 0.10 related to this cost for the 2-carrier MaBS, 3-carrier MaBS, MiBS, PBS and Wi-Fi BS. Consequently in this paper, we assume 20 k€ OPEX for the new 3-carrier MaBS site.

TABLE II. CAPEX, OPEX AND RESULTING DISCOUNTED COST ESTIMATES PER BASE STATION CLASS FOR GREENFIELD DEPLOYMENT (ALL AMOUNTS IN [K€]).

BS Class/ LTE-A and IEEE 802.11ac	Initial CAPEX (Investment)	Annual OPEX	Total discounted cost in period of 10 years
Macro (1 carrier)	72.9	15.5	152.67
Macro (2 carriers)	96.2	17.8	186.47
Macro (3 carriers)	120.0	20.0	220.15
Micro	35.8	10.4	90.73
Pico	13.5	3.4	31.26
Femto	1.0	0.5	3.72
Wi- Fi	2.5	1.6	12.17

According to Markendahl and Mäkitalo [3], we assume 10 k€ for the existing site. For the FBS, Markendahl and Mäkitalo in [3] estimates the annual operational cost to be 0.5 k€ per BS. To summarize the discussion on cost estimates, Table II outlines the resulting discounted cost per the considered newly deployed BS class as calculated according to (3).

B. Spectrum Cost Analysis

Alternatively, and if possible, MNOs could increase the number of carriers by adding additional spectrum, which could replace the deployment of new sites. This brings spectrum to an essential asset as it could be a substitute for new sites.

The more bandwidth that can be used at one site the higher the capacity. Currently, across Europe MNOs have licensed spectrum at different bands and the carriers in between are set at 800 MHz, 900 MHz, 1800 MHz, 2100 MHz and 2600 MHz bands. All parts from the available bandwidth provide different performance for coverage and capacity.

MNOs are annualizing the CAPEX related to the spectrum licenses for the period of their validity, which is mostly 10 years and usually no more than 20 years. Additionally MNOs have OPEX per MHz related to the annual frequency charges.

From today's perspective and according to the ongoing developments in the European telecommunication markets, most of the used spectrum is amortized, excluding the part of the spectrum from 790 MHz to 862 MHz (so called Digital Dividend - DD) that was acquired by the MNOs in the past few years in most of the European countries. Based on the benchmark analysis of the data collected from the European National Regulatory Authorities websites [26], the average annual frequency fee per MHz is below 1 EUR/MHz and population and maximum 10 EUR/MHz and km². Furthermore, according to BEREC [26], the Figure 1 depicts the invested price in DD band per MHz and per km² (Note that, for Netherlands and United Kingdom the price is 1525 and 736 EUR/MHz/km².

According to PWC [27], the invested price in DD band per MHz and per population moves from 0.2 EUR in Croatia up to 0.8 EUR in Italy.

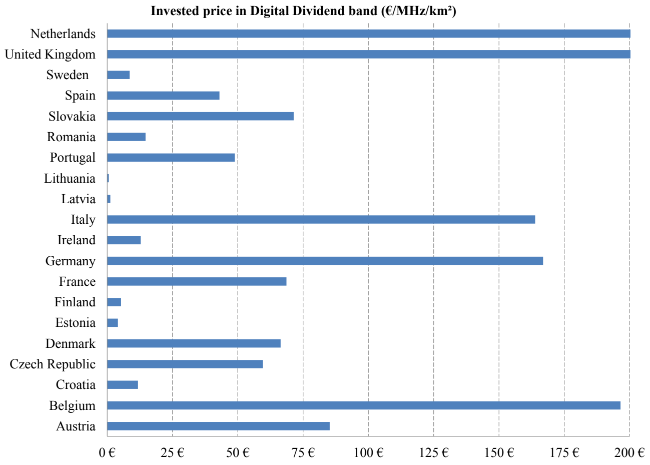


Fig. 1. Benchmark of the mean price (in €/paired MHz/km²) paid by the MNOs of the European Union Member States in the Digital Dividend (792-862 MHz) spectrum auctions.

IV. Investment Case Study

A. Case Study Description

According to Johansson et al. [2], the different BSs will minimize cost for different scenarios. Nevertheless, for the sake of simplicity, first we will perform cost modeling through case study of different base stations separately. Based on the results of separate analysis, than we will dimension the network of the analyzed service area as a combination of a of a macro layer solution, using existing sites and as much available spectrum as possible, with a supporting small cells network. Accordingly, based on the per unit cost estimates from Table I and Table II in this section, we will assess how the total investment cost (initial CAPEX) of the wireless network deployed within the particular area, varies as a function of the user demand. Furthermore, we will apply different deployment scenario combining the amount of the bandwidth and BS classes. In particular, we consider building of the new office center in the 1 km² urban indoor area through construction of ten 5 floor buildings hosting 10.000 workers. Consequently, we will not analyze the MiBS and PBS options out of the small cell deployments, but only the strict indoor solution of small cells represented by FBS alongside with the Wi-Fi. For the macro layer, we will consider the CAPEX needed for deployment of three-sector MaBS supporting three frequency carriers. Nevertheless, for the capacity estimates, we will consider that only single carrier is in use, to make the comparison between BS types simple.

In line with Figure 1, the cost of 1 MHz per the system area of 1 km² is negligible compared to the cost of even single carrier MaBS. Consequently, we will ignore the CAPEX inputs for the spectrum within the estimation of the total investment costs calculated according to (4).

TABLE III. Conversion of Load/User/Month to the User Data Rates (Mbps) and Capacity per Area Unit (Gbps/km²).

Demand	GB/user/month	Mbps/user	Gbps/km²
Moderate	44.0	0.407	4.0
High	110.0	1.019	10.0

B. Traffic Demand

Based on [28], in 2013 around 95% of the total global mobile traffic was generated by smartphones (62%), laptops (24.5%) and tablets (8.5%) with around 0.5 GB/month from smartphone user and 2.6 times more from the laptop users and 4.6 times more from tablets (only 3% of the users generated more than 5 GB/month and 24% more than 2 GB/month). The same source predicts that the average usage per month of smartphones will rise x 5 times (up to 2.7 GB) by 2018 having 66% from the total traffic and that tablet share will be more than 18%. Following the same ratios, we could draw conclusion that the average usage per month in 2018 will be around 12.2 GB and 6.9 GB for tablets and laptops respectively. Furthermore, [29] predicts an average N. American mobile user to consume 6 GB/month in 2017.

Consequently, in order to ensure future-proof network (e.g., beyond 2020), we will perform the dimensioning of the network from our case study with the following two demand levels: moderate demand or in average 44 GB/user/month and high demand of 110 GB/user/month.

We consider that the usage will be spread out over 8 hours per day, translating into a busy hour rate of 12.5%, in line with the industry standard [30]. Conversion of the load/user/month to the user data rates (Mbps) and capacity per area unit for 10 000 users (Gbps/km²) is given in Table III, for the 8 busy hours. In this paper, we consider uniform traffic distribution within the considered area.

C. Macro cellular Deployments

Assuming the spectral efficiency of 3.8 bit/s/Hz/cell of outdoor LTE-A RAT, the achieved capacity with a single carrier three-sector MaBS site is 114.0 Mbps, 228.0 Mbps and 342.0 Mbps with 10 MHz, 20 MHz and 30 MHz of spectrum, respectively (calculated in line with (2)).

1) Initial Scenario

Since a cell area of 1 km² corresponds to a cell radius of 0.57 km (according to (1)), our requirements on average user data rates during busy hours would be met even at the cell borders with the high broadband demand (~ 1.0 Mbps what is in line with the data rate of 1.0 Mbps as assumed in [11]).

Within the initial scenario, we perform the cost-capacity analysis using 20 MHz for the macro-layer in the 2.6 GHz band with the average spectral efficiency of LTE-A RAT. In accordance with [3], for the MaBS site re-use scenario, we estimate the total CAPEX of 20 k€ for existing site (the cost needed to upgrade an existing site is estimated to 10 k€ and the cost for radio equipment supporting three sectors and 5–20 MHz to 10 k€). Based on Tables I and II, Table IV summarizes the total invested costs for the moderate and high demand estimates. It is noticeable that the investment to satisfy the high demand with the implementation of the existing MaBS is almost half than the cost needed to ensure 2.5 times less capacity with new MaBS sites.

2) Wall Penetration Losses Compensation Scenario

When trying to compensate for the wall penetration losses, two options are possible according to Markendahl and Mäkitalo [3] and Markendahl [4]: building a denser 2.6 GHz network and deployment using 10 MHz within the 800 MHz band, i.e., better indoor coverage.

TABLE IV. INVESTMENTS AND CAPACITY (MACRO SITES INITIAL DEPLOYMENT - CASE 1).

Macro Initial Scenario (2.6 GHz)		Number of sites	Total CAPEX M€	Capacity (Gbps)
Site	*Demand*			
New	Moderate	18	2.16	4.1
New	High	44	5.3	10.03
Reuse	Moderate	18	0.36	4.1
Reuse	High	44	0.88	10.03

Markendahl and Mäkitalo in [3] calculated that in order to compensate the additional 12 dB of attenuation (the difference between operation in the 800 MHz and the 2.6 GHz band), 5 time denser network should build at 2.6 GHz band. Consequently, Table V summarizes the cost-capacity outcomes of this scenario. We can see that in order to compensate the wall penetration losses with the MaBS solution, the deployment of a large number of new sites is very costly. Again, the re-use of existing sites leads to less costly deployment even when many sites need to be equipped with new radio transceivers for the high demand. Still, due to the high coverage performance, the most cost-efficient option in case of high demand is the reuse of the existing sites with 10 MHz in 800 MHz band.

3) Carrier Aggregation Scenario

According to Qualcomm [31], carrier aggregation as characteristic of LTE-A RAT, allows combining lower and higher bands — leveraging better coverage of the former with higher availability of the latter (up to 5 carriers and up to 100 MHz supported in standards). In order to fully assess the cost-efficiency possibilities, we create one more deployment scenario assuming the aggregation of the both frequency carriers at 800 MHz and 2.6 GHz bands. By this the bandwidth available will be increased to 30 MHz, and exactly this is going to be the solution how to increase the capacity (even for 3 times) compared to the use of only 10 MHz bandwidth in 800 MHz band, but without increase the number of sites due to coverage reasons as in the case with 2.6 GHz deployment. From the cost perspective, this will mean that we need to install two type of different radio equipment per BS. Consequently, the CAPEX will increase for additional 10 k €, and the total CAPEX per site will be 130 k€ for the new sites and 30k € for the existing sites. According to pervious estimations, we will assume OPEX of 20 k€ for new and 10 k€ for the existing MaBSs.

The number of needed BS sites using carrier aggregation functionality and the relevant costs-capacity outcomes are summarized in Table VI.

TABLE V. INVESTMENTS AND CAPACITY (MACRO SITES WALL LOSSES COMPENSATION DEPLOYMENT - CASE 2).

Macro Wall Losses Compensat. (0.8 or 2.6 GHz)		Number of sites	Total CAPEX M€	Capac. (Gbps)
Site	*Demand*			
New 0.8 GHz	Mod.	36	4.32	4.1
New 0.8 GHz	High.	88	10.56	10.03
Reuse 0.8 GHz	Mod.	36	0.72	4.1
Reuse 0.8 GHz	High.	88	1.76	10.03
New 5 x 2.6 GHz	Mod.	90	10.8	20.5
New 5 x 2.6 GHz	High.	220	26.4	50.16
Reuse 5 x 2.6 GHz	Mod.	90	1.8	20.5
Reuse 5 x 2.6 GHz	High.	220	4.4	50.16

TABLE VI. INVESTMENTS AND CAPACITY (MACRO SITES WITH CARRIER AGGREGATION - CASE 3).

Macro Carr. Aggr. (0.8 & 2.6 GHz)		Number of sites	Total CAPEX M€	Capacity (Gbps)
Site	*Demand*			
New	Moderate	12	1.56	4.1
New	High	30	3.9	10.26
Reuse	Moderate	12	0.36	4.1
Reuse	High	30	0.9	10.26

Findings show that for around 0.9 M€ needed to upgrade the existing sites, the high user demand will be ensured. Further, with 1.56 M€ of investment and construction of new sites the high demands can be satisfied, too.

D. Femto Cell and Wi-Fi Deployments

In line with [3], and explanations for the maximum numbers of users per access point for FBS and Wi-Fi given in Section II above, we consider different options of the user oriented and coverage oriented approaches. Since the construction of the new office center is green-field, we will assume that previously there were no small cell installations within the considered area of 1 km². Consequently, the Table VII summarizes the cost-capacity figures for the FBS and Wi-Fi deployments. As expected with any of the considered scenarios of FBS and Wi-Fi, the provisioning of the demanded capacity will be achieved. The coverage is main cost driver for these two scenarios and high density indicates high network costs.

V. COMPARATIVE DISCUSSION RELATED TO THE SEPARATE NETWORK DEPLOYMENTS

Assuming the total investment budget of 3.0 Million € (M €), we compare in Figure 2 the investment costs in M € for separate network deployment scenarios as function of user demand in Gbps. It is noticeable that LTE-A MaBS deployment with site re-use and carrier aggregation in place, has the lowest cost for the capacities below 2.0 Gbps. Even LTE-A MaBS deployment with new sites and carrier aggregation in place is more cost effective option compared to the Macro 5xtime denser deployment and site reuse at 2.6 GHz band. Hence, the LTE-A RAT and carrier aggregation functionality from cost perspective could be acceptable MaBS deployment scenario for the new market entrant as well, since with it the new comer will be able to achieve comparable profitability with the existing operators for relatively high demand levels.

TABLE VII. INVESTMENTS AND CAPACITY (FBS LTE-A BASED AND WI-FI IEEE 802.11AC DEPLOYMENTS).

Femto Cells and Wi-Fi	No. of sites		CAPEX M€		Capac. (Gbps)	
	FBS	Wi-Fi	FBS	Wi-Fi	FBS	Wi-Fi
4 users / BS	2500	2500	2.5	6.25	330	3250
8 users / BS	1250	1250	1.25	3.13	165	1625
16 users / BS	625	625	0.63	1.56	82.5	812.5
32 users / BS	313	313	0.32	0.78	41.3	406.9
4 BS / floor	200	200	0.2	0.5	26.4	260
8 BS / floor	400	400	0.4	1.0	52.8	520
16 BS / floor	800	800	0.8	2.0	105.6	1040
32 BS / floor	1600	1600	1.6	4.00	211.2	2080

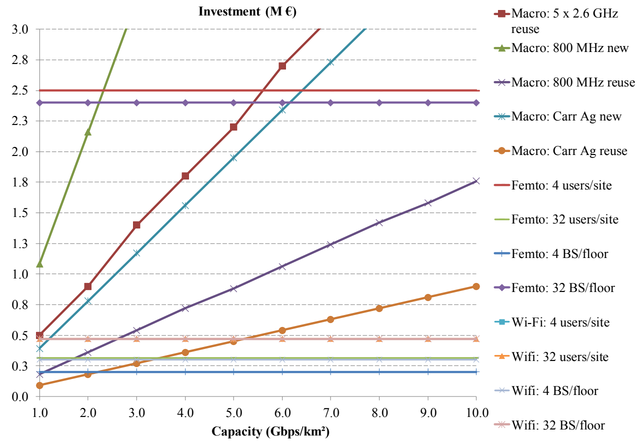


Fig. 2. Comparison of macro and small cell deployment costs as function of the user demand, with the LTE-A and IEEE 802.11ac, respectively.

From other side, deployment with the reuse of the existing MaBS with 10 MHz spectrum in the 800 MHz band causes achieving high demand with tolerable investment of 1,75 M€ due to the superb coverage and penetration performance of the 800 MHz carrier frequency. For the existing mobile operator missing spectrum in the 800 MHz, an option will be to reuse existing sites with 5 time higher density, what is more cost-effective solution than MaBs deployment with new sites in the 800 MHz band what in fact is the less cost efficient option. Thus, we can draw a conclusion that it is very important if new MaBS sites need to be deployed or not. In general, for the MaBS deployments it could be noticed that the slope of the lines depends on the number of sites that are needed and especially if new sites need to be deployed. The performances of FBS and Wi-Fi are different. As we already considered those types of indoor deployments are coverage, rather than capacity limited. Their cost depend form the density of BS used. As shown in Figure 2, for dense network deployments 4 users per FBS/Wi-Fi or 32 FBS/Wi-Fi sites per floor, is less cost-effective option comparing to most of the MaBS deployments unless the user demand is extremely high (above 6.5 Gbps). FBS/Wi-Fi deployments are cost-efficient when single site can support higher number of users (e.g., 32 per site or 4 sites per floor).

Thus, for the capacities above 2.0 Gbps, the most cost-effective deployment option is the utilization of 4 FBS per floor. A comparison of FBS and Wi-Fi shows that the FBS solution is more cost effective than Wi-Fi deployment, but from the capacity long-term perspective the better option should be IEEE 802.11ac Wi-Fi deployment due to its superb capacity performance.

VI. Demonstration of the Combined Cost-Capacity Modeling

In the previous section, we have conducted the cost-capacity modeling through case study of different BSs separately and focusing only at the investment performed within the first year. Following the findings of the separate solutions, here, we will demonstrate the network dimensioning of the analyzed service area as a combination of most cost-effective macro and small cell or Wi-Fi solutions.

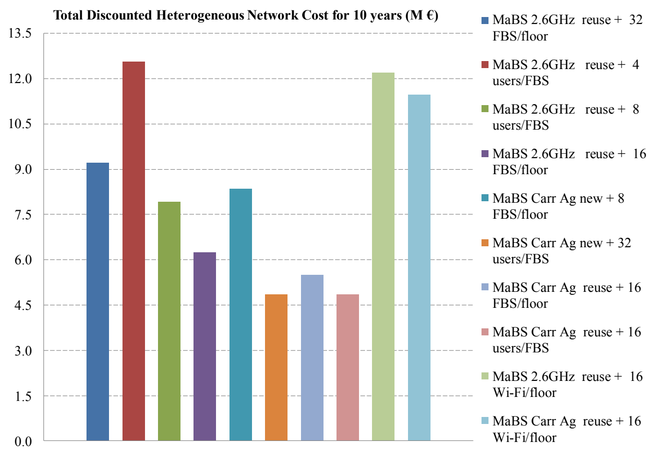


Fig. 3. Wireless heterogeneous network total discounted cost for the period of 10 years, jointly deployed by categories to satisfy high demand level of 10 Gbps/km ² with the LTE-A and IEEE 802.11ac.

This could be as of particular interest of MNO having existing network deployment within the analyzed area. Thus, in the spotlight once again comes the initial scenario with the usage of 20 MHz in the 2.6 GHz band for the macro layer, identified as insufficient to compensate the wall penetration losses arising with the construction of the office center.

The graphical representation of the total discounted cost for various heterogeneous network deployments in 10 years period is shown in Figure 3. The results are yield in accordance with (3) and (4), the total discounted cost estimates per different BS classes (CAPEX + OPEX in present value for the period of 10 years and WACC = 12%) and findings for the number of BS (as per Tables IV – VII) needed to satisfy the high demand level of 10 Gbps/km ².

As some of the FBS and Wi-Fi options produce capacity overprovisioning (e.g., 4-8 user per BS or 16-32 BS per floor), we combine some of those deployments only with the initial MaBS scenario. The rest of the FBS and W-Fi solutions are combined with the MaBS scenarios which ensure wall penetration losses compensation, too.

It could be noticed that MNO having deployed macro network with 20 MHz in the 2.6 GHz network, instead of investing in additional spectrum or deploying denser network, it could compensate the indoor wall penetration losses by deploying 16 FBS sites per floor. That total discounted cost-efficient level of around 6.0 M€ is comparable for instance with deployment of new MaBS sites with carrier aggregation and 32 users per FBS indoor deployment what in fact is the most cost efficient combined macro/small cell deployment for still acceptable QoS from capacity perspective.

VII. Conclusion

We introduced a model for evaluation of the total deployment costs of heterogeneous wireless access networks. The model uses up to date inputs of the unit cost of particular base station class which is characterized with specific coverage and capacity parameters. For the cellular deployments, we use the forthcoming LTE-A RAT and for the WLAN networks, we consider the future-proof IEEE 802.11ac standard.

Through the investment case study, which considers construction of large office center, we have compared the cost-capacity performance for macro and small cell deployments as a function of moderate to very high user demand levels. The study analyzed deployments in both the 800 MHz and 2.6 GHz bands as well as the scenario of aggregated carriers in these bands. Findings show that the macro cell deployment scenarios show linear increase with demand. In order to satisfy moderate demand levels, it can be concluded that the re-use of sites, have a large impact also when a "denser" macro network is deployed in order to compensate for wall attenuation. The re-use of the existing macro sites with the low-end frequency carriers at 800 MHz, represents moderate cost-efficiency compared to other solutions.

Still, the solution to deploy the denser network at 2.6 GHz band with re-use of the existing sites is more cost-efficient than the solution to construct new sites with 800 MHz carrier, what shows the importance of the spectrum available, too. Hence, the key finding is that use of carrier aggregation functionality of LTE-A will significantly increase the cost-effectiveness of the macrocellular deployment. Thus, with enabling aggregation of the carriers in the band of 800 MHz and of 2.6 GHz on the existing sites, we create the most cost-efficient deployment for moderate demand levels.

On the other side, the indoor deployed femto cell and Wi-Fi solutions (being only coverage limited) are most cost efficient only for the higher to extreme user demands. Results indicate that FBS/Wi-Fi significantly become cost-efficient when single site can support higher number of users, basically due to the very low unit cost compared to the equipment cost of the higher order cellular deployments. With regard to the joint heterogeneous deployment, we determine that for operator holding less spectrum and in the upper bands, instead of investing in additional spectrum or deploying denser network, it could compensate the indoor wall penetration losses by deploying the acceptable number of FBS sites per floor from perspective of high demand levels and taking into account the "time value of money".

Further studies in this field could investigate the cooperative layouts of macro with femto cells or Wi-Fi by consideration of the beyond 2020 mobile and wireless system targets [22].

REFERENCES

[1] K. Johansson, "Cost Effective Deployment Strategies for Heterogeneous Wireless Networks", PhD Dissertation. The Royal Institute of Technology, Stockholm, 2007.

[2] K. Johansson, A. Furuskar, P. Karlsson, and J. Zander, "Relation between base station characteristics and cost structure in cellular systems In Personal, Indoor and Mobile Radio Communications," PIMRC 2004. 15th IEEE International Symposium on, volume 4, May 2004, pp. 2627– 2631, doi: 10.1109/PIMRC.2004.1368795.

[3] J. Markendahl and Ö. Mäkitalo, "A comparative study of deployment options, capacity and cost structure for macrocellular and femtocell networks Personal, Indoor and Mobile Radio Communications Workshops (PIMRC Workshops), 2010 IEEE 21st International Symposium, Istanbul, Sep. 2010, pp. 145-150, doi: 10.1109/PIMRCW.2010.5670351.

[4] J. Markendahl, "Mobile Network Operators and Cooperation", PhD Dissertation. The Royal Institute of Technology, Stockholm, 2011.

[5] Z. Frias and J. Pérez, "Techno-economic analysis of femtocell deployment in long-term evolution networks," EURASIP Journal on Wireless Communications and Networking, volume: 2012, 2012, issue: 1 pp. 288-302.

[6] ETSI TR 136 913 V10.0.0 (2011-04) LTE: ETSI, 2011.

[7] ETSI TR 136 912 V11.0.0 (2012-10): ETSI(2012).

[8] IEEE 802.11acTM-2013. IEEE Standards Association 2014.

[9] Analysis Mason,"Wireless network traffic 2010–2015", 2010.

[10] K. Johansson, J. Zander, and A. Furuskär, "Modelling the cost of heterogeneous wireless access networks". Int. J. Mobile Network Design and Innovation, Special Issue on Planning and Optimisation of Wireless Networks, vol. 2, no. 1, 2007, May 2007.

[11] H. Holma and A. Toskala, (ed), LTE for UMTS – OFDMA and SC-FDMA Based Radio Access. John Wiley & Sons, 2009.

[12] K. Johansson and A. Furuskär, "Cost efficient capacity expansion strategies using multi-access networks," Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st, Volume 5, Jun. 2005, pp. 2989 – 2993, doi: 10.1109/VETECS.2005.1543895.

[13] S. Choi, "Femtocell vs. WiFi". The 22nd High-Speed Network Workshop. Multimedia & Wireless Networking Lab, 2012.

[14] B. G. Mölleryd, J. Markendahl, and Ö. Mäkitalo, "Spectrum valuation derived from network deployment and strategic positioning with different levels of spectrum in 800 MHz," In Proceedings of 8th Biennial and Silver Anniversary ITS Conference, Tokyo, Jun. 2010.

[15] REPORT ITU-R M.2134, Requirements related to technical performance for IMT-Advanced radio interface(s), 2008.

[16] R4-071231, Open and Closed Access for Home NodeBs. "Nortel, Vodafone", 3GPP TSG RAN Working Group 4 meeting #44, 2007.

[17] Femto Forum, Interference Management in UMTS Femtocells, Feb. 2010, Retrived: May 30, 2014 from http://www.smallcellforum.org/.

[18] M. Gast, 802.11 Wireless Networks – The Definitive Guide. 2nd ed. O'Reilly, 2005.

[19] Y. Xiao, "IEEE 802.11n: Enhancements for higher throughput in wireless LANs," Wireless Communications, IEEE (Volume: 12, Issue: 6), Dec. 2005, pp. 82 - 91, doi: 10.1109/MWC.2005.1561948.

[20] C. Wang and H. Wei, "IEEE 802.11n MAC Enhancement and Performance Evaluation," Mobile Networks and Applications volume 14, 6, 2009, pp. 760-771.

[21] White Paper. 802.11ac: The 5th Generation of Wi-Fi. Cisco, 2014.

[22] METIS (Mobile and wireless communications Enablers for the Twenty-twenty Information Society) Project, "Scenarios, requirements and KPIs for 5G mobile and wireless system", Document Number: ICT-317669-METIS/D1.1, 2013.

[23] Asus, RT-AC68U AC1900 Wi-Fi Router specs. PS WORLD, Retrived: May 30, 2014 from http://www.pcworld.com.

[24] White Paper. Proven-Carier Grade Wi-Fi Solutions. Motorola, 2011.

[25] White Paper. Carrier Class – The Esential Ingridient for Succesfull Metro Wi-Fi. Zhone, 2008.

[26] BEREC, Member and Observer National Regulatory Authorities, Retrived: May 30, 2014 from http://berec.europa.eu/eng/links/.

[27] PWC, "Digital dividend in Southeast Europe", November 2012.

[28] Cisco, "Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018", Feb. 2014.

[29] Cisco. Average N. American mobile user to consume 6 GB/month in 2017. Fierce Wireless, Feb. 2013, Retrived: May 30, 2014 from http://www.fiercewireless.com/story/cisco-average-n-american-mobile-user-consume-6-gbmonth-2017/2013-02-05.

[30] A. Furuskär, M. Almgren, and K. Johansson, "An Infrastructure Cost Evaluation of Single- and Multi-Access Networks with Heterogeneous Traffic Density," IEEE Vehicular Technology Conference, Vol. 5, Jun. 2005, pp. 3166 – 3170, doi: 10.1109/MWC.2005.1561948.

[31] Qualcomm, LTE Advanced—Leading in chipsets and evolution, Aug. 2013, Retrived: May 30, 2014 from http://www.4gmast.nl/Upload/wireless-networks-lte_advanced-leading_in_chipsets_and_evolution.v6.20130827.pdf

Interference Modelling and Analysis of Random FDMA schemes in Ultra Narrowband Networks

Minh-Tien Do

Sigfox Wireless Company
Labège, France
Email: minhtien.do@sigfox.com

Claire Goursaud and Jean-Marie Gorce

INSA-Lyon, CITI-Lab
Lyon, France
Email: claire.goursaud@insa-lyon.fr
jean-marie.gorce@insa-lyon.fr

Abstract—**Ultra narrow band (UNB) transmission is a very promising technology for low-throughput wireless sensor networks. This technology has already been deployed and has proved to be ultra-efficient for point-to-point communications in terms of power-efficiency, and coverage area. This paper introduces this technology and gives some insights on the scalability of UNB for a multi-point to point network. In particular, we present a new multiple access scheme: random frequency division multiple access (R-FDMA) and study the impact of the induced interference on the system performance in terms of bit error rate and outage probability. To this aim, we propose and design a simplified model to describe the interference impact. Thanks to this model, we theoretically derive BER and OP expressions for the lower, approximated and upper case. This enables us to evaluate the performance capacity, by determining the maximum number of simultaneous users that can be served.**

Keywords–Wireless sensors networks; M2M/IoT applications; Random FDMA; Aggregate interference; Ultra narrow band.

I. INTRODUCTION

In current trend, the internet of things (IoTs) and wireless sensor networks (WSNs) share many common constraints [1], and thus, the communication techniques applied for WSNs could be reused for IoTs and machine-to-machine (M2M). The challenges are the connection of countless wireless devices and the requirement of cost-effective, power-efficient and scalable network. In networks for applications, such as temperature monitoring, electrical metering etc., nodes send dynamically a small amount of data. As a consequence, a high bit rate is not mandatory for each link. Therefore, ultra narrow-band (UNB) transmissions can be used for such low-throughput networks.

UNB consists of sending the information occupying a very narrow frequency band with the binary-phase-shift-keying (BPSK) modulation. The BPSK modulation is used because it satisfies power-efficiency, bandwidth-efficiency and cost-effectiveness for low-throughput network in long range communication [2]. Besides, as the occupied band is reduced, the noise contribution is lessen at the receiver. Consequently, for a given targeted error probability, the reception power sensitivity is very low, enabling a very large coverage area using a single base-station (more than 50 km in open field).

With such an extended coverage, a large amount of source nodes are eligible to be served and will compete for transmission. Thus, the medium access control (MAC) protocol is important to consider. The contention-free channel access methods are not efficient with respect to the low quantity of information to be transferred and would lead to a waste of time for protocols or synchronization issues. Contrarily, the random access protocols are a promising solution, as they present more flexibility to manage bursty and random transmissions.

As verified in [3][4], most of the MAC studies consider that the nodes share the same frequency channel, and focus on the decision of the moment to transmit. Nonetheless, studies on the multi-channel MAC also consider the frequency as a random variable [5][6][7]. However, these studies consider predefined disjoint channels, which is not a realistic assumption in UNB networks. Indeed, at typical transmission frequency 800 MHz, and a typical oscillation jitter 0.5 ppm - 2 ppm, there is an uncertainty on the frequency positioning is around 400 Hz, which is bigger than the transmission band. As a consequence, with UNB technology, random frequency multiple access (R-FDMA) scheme has to be considered, as we proved in [8]. The network behaves as if each node transmitted in a bursty way to access to the medium, and at a frequency chosen randomly in the available bandwidth.

Consequently, at the PHY layer, besides the effect of classical channel impairments such as fading, shadowing effect, inter-symbol interference and noise [9][10], in R-FDMA scheme, the system performance depends also on the carrier frequency distribution and the corresponding interference term resulting from physical channels overlap. While the performance of the single link is easy to obtain, no accurate model for multiple links has been proposed. Specifically, the behavior of the interference induced by a large number of unconstrained nodes (both in time and frequency) over a wide area around the sink has not yet been studied. For certain classes of node distribution, most notably Poisson point processes, and attenuation laws, closed-form results are available for both interference term and signal-to-interference ratios (SIR), which determine the network performance [11][12][13][14][15]. However, as in MAC studies, the users are either transmitting in the same channel (i.e., with the same carrier frequency: thus highly interfering), or in adjacent channels (thus barely interfering). But, in the case of continuous R-FDMA, the frequencies are selected in a continuous way in the total band and lead potentially to all values, independently of the path-loss. Therefore, a new analysis of the system performance needs to be done, to take into account this new specificity.

In this paper, we propose to study the interference of Random FDMA schemes in UNB network. We characterize the system performance by understanding and modeling the distribution of the aggregate interference power (AIP). The others channel impairments are neglected. We propose an approximation for the AIP and derive a closed-form of the probability density function (PDF) of channel interference. This enables us to provide an upper and lower bound beyond to estimate the system performance.

The rest of the paper is organized as follows. Section II

ISBN: 978-1-61208-360-5

presents the wireless network model for UNB and describes the considered R-FDMA scheme. In Section III, we present the theoretical interference analysis and simplified models that are used in next section for the system performance evaluation. Then, the estimated capacity network using the simplified models are presented in Section IV. Finally, we conclude in Section V.

II. Transmission Model

A. Ultra Narrow Band Transmission

UNB refers to the fact that the individual bands used at the transmission sides are very narrow compared to the whole available bandwidth (typically 1:100). While digital or analog data of narrow band radio system are transmitted and received over a few kHz [9], UNB signals require around 100 Hz only, which can be achieved with highly selective FIR filters. Such transmissions have several benefits: flat fading can be assumed, which highly simplifies the system analysis and the receiver, while a higher number of users can be supported.

UNB technology is currently deployed, e.g., in Sigfox's networks [16]. In these deployments, a star topology is used, where base-stations centered in large cells receive the data from a huge amount of source nodes spread over. Because of the ultra narrow spectral occupation, the noise contribution is very low (around -150 dBm at $T = 290$ K). So, contrary to classical deployments, such technology enables an exceptionally large-scale wireless connection thanks to the ability to successfully demodulate an extremely low received power signal (-142 dBm). These advantages allow data transmission in highly constrained environments where former technologies cannot operate and a possibility to cover a very large area with a very small number of base stations, reducing network management and deployment fees of several orders of magnitude.

B. R-FDMA Scheme Definition

In a random access frequency network, four main problems must be considered: the asynchronicity access of node in the wireless medium, randomness both in time and frequency domain and lack of contention based protocols. To illustrate the system behavior, a toy-example is schematized in Fig.1. It represents the time and frequency use of the channel for 4 active users.

The *randomness in time domain* has an impact on the number of users N that will be active at the same time. This value depends on several parameters: the number of possible users in the cell, the length (in time) of the packet to transmit, and the periodicity of the transmission. We present our results as a function of $k = N - 1$ the number of interfering users.

Furthermore, the *asynchronicity* permits to suppress the traffic overload needed for synchronization, but leads to varying interference levels during the transmission of a given packet, as packets do not start (and stop) at the same time. In order to simplify the analysis discussed in this work, we will not evaluate the performance evolution during the whole packet transmission, but only at a given point in time. For example, in Fig.1, at $t = t_0$ only 3 users among the 4 users are transmitting.

The *randomness in frequency domain* has an impact on the position of each active users carrier in the total band. Thus, it affects the interference suffered by a given user, which depends on the spacing δ_f between the users carrier frequency and the interferers one. The Random FDMA schemes could be divided into two kinds of frequency randomness [8] continuous and discrete. In the discrete case, the carriers are chosen at random in a discrete and pre-defined subset of frequencies. But, in order to take into account the carrier imprecision due to the jitter, we consider only continuous random frequency division multiple access, where the carriers can be chosen at random in the continuous available frequency band. In this case, from the receiver point of view (i.e., on base-station side), the monitored bandwidth is filled from time to time with a set of signals of interest occupying a small amount of total spectrum and centered around unpredictable carrier frequencies. Thus, in order to handle demodulation, efficient software defined radio algorithms have been designed to analyze the total band, determine transmitter activity and retrieve data they are transmitting. These algorithms are currently deployed in SigFoxs network, and do not fall in the scope of this paper.

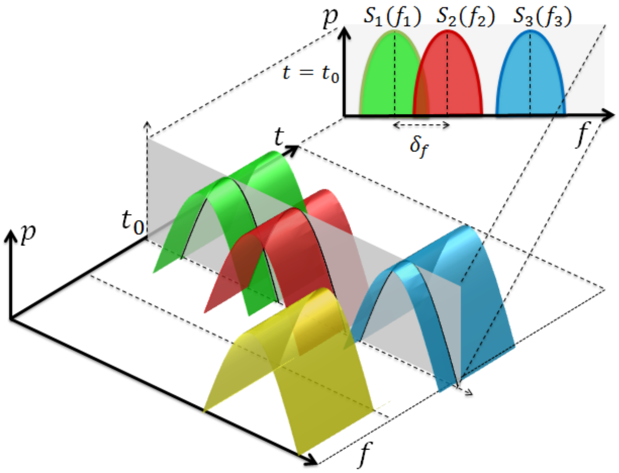


Figure 1: Example of temporal & spectral repartition of users.

The *lack of contention based protocols* implies that each user is transmitting without any knowledge of carrier frequencies being used in the cell. Thus, this induces interference (when at least 2 users are transmitting at the same moment and there is an overlap between the individual transmission bands). For example, in Fig.1, the green user starts transmitting even if the red one is already using the band in common.

Furthermore, we should note that R-FDMA allows the use of transmitters whose frequency is unconstrained (except for being in the transmission bandwidth). In practice, the randomness in frequency domain is easily done: each node has its own transmission frequency, which it not controlled by the network, but defined by the node components (electrical components an oscillator jitter), and may vary naturally (depending on different parameters such as temperature and age of the device). Thus, factory constraints are relaxed, and the network will not be sensible to temperature variations and other environmental parameters that can affect the carrier. Thus, cheaper nodes can be used.

As a consequence, R-FDMA is promising for smart metering where a massive amount of devices have to be connected to the Internet, provided that the randomness does not highly degrade the performances.

C. System Mathematical Model And Parameters

As described in the previous section, the main characteristic of the considered network using R-FDMA at a given point of time is that each active user is transmitting at a carrier frequency randomly chosen in a given band. As a consequence, interference contribution is non-controlled and can lead to transmission errors. Consider a multiple access channel with $N = k + 1$ active transmitters (note that N is much smaller than the number of nodes that are actually in the cell). The total received signal at the base-station can be expressed as:

$$r(t) = \sum_{i=1}^{k+1} s_i(t) \cdot g(f_i, t) \otimes h_i(t) + n(t) \qquad (1)$$

where $s_i(t), \forall i \in [1, \dots, k+1]$ are the BPSK symbols sent by the active user i, $g(f_i, t)$ the impulse response of the emission FIR filter (centered at f_i); $h_i(t)$ is the path-loss of the corresponding link, and $n(t)$ is an additive white Gaussian noise with zero mean, and whose variance is σ^2.

For the sake of simplicity in this analysis, we consider that $h_i(t) = \delta(t)$, $\forall i \in [1, \dots, k+1]$. This corresponds to the worst case where all users are at the same distance of the base station and experience the same flat channel. At the base station, the received signal is analyzed to track possible transmissions in the total band (BW), and filtered at the desired frequency. Without loss of generality, we consider in this paper that the desired user is #1. The signal used for data recovery is thus:

$$\begin{aligned} r^{'}(t) &= r(t) \otimes g(f_1, t) && (2) \\ &= \sum_{i=1}^{k+1} s_i(t) \cdot g(f_i, t) \otimes g(f_1, t) + n(t) \otimes g(f_1, t) && (3) \end{aligned}$$

To evaluate the system performances, we use the signal to interference plus noise ratio (SINR), which is expressed as:

$$\text{SINR} = \frac{P_s}{N_{tot} + P_I} \qquad (4)$$

where P_s is the received power of the desired user, P_I the aggregate interference, and N_{tot} the noise contribution. These powers are estimated at a given time, and normalized with respect to $P_s = |G(f_1, t)|^2$ with $G(f_1, t)$ the frequency response of the FIR filter. The value of P_I depends on the spacing between the carriers frequency, and its estimation will be described in the next section. We deduce the bit error rate (BER) of the BPSK transmission from the SINR as follow:

$$\text{BER(SINR)} = Q(\sqrt{SINR}) \qquad (5)$$

A data transmission is considered successful if the received BER is below a predefined threshold $\beta = 10^{-3}$, otherwise, the data are considered lost. Thus, we consider the outage probability (OP) being expressed:

$$Pr(OP) = Pr(BER \geq \beta) = Pr(BER \geq 10^{-3}) \qquad (6)$$

The simulation results shown in Section III and IV, the BER and OP are obtained with respect to (5), (6) (with a noise power 100 dB under the signal of interest).

III. Theoretical Interference Analysis

As described in Section II, the R-FDMA scheme solves a waste of communication resources for WSNs where the users send a short message. However, it leads to interference that must be quantified. Therefore, the goal of this study is to analyze the aggregated interference power (AIP) and propose the simplified model for UNB network based on R-FDMA scheme.

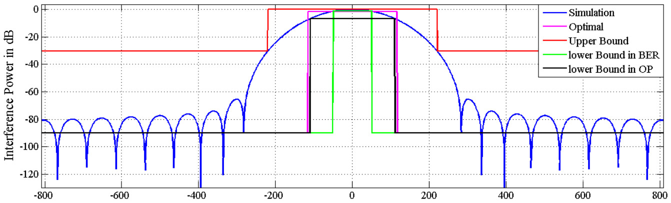


Figure 2: Behavior of interference vs frequency difference δ_f.

A. Modelization of a single interferer contribution

In the single interferer case, we consider the interference power created by a unique interferer. We assume that there are only $N = 2$ active users using R-FDMA scheme (i.e., the useful signal and $k = 1$ interfering signal). The interference power can be derived at a given time by multiplying the frequency responses of the useful signal and interfering signal.

$$P_I(t) = \mid G(f_1, t) \cdot G(f_2, t) \mid \qquad (7)$$

In (7), the only parameter that will influence $P_I(t)$ is the relative frequency positioning $\delta_f = |f_1 - f_2|$ between the carriers used by the active users. Therefore, we model the interference level as a function of the frequency shift between the 2 active users $\delta_f = |f_1 - f_2|$:

$$P_I(t) = \mid G(f_1, t) \cdot G(f_2, t) \mid = P(\delta_f, t) \qquad (8)$$

From now on, as we focus on the interference at a given sample time normalized to P_s, we neglect the time variable in the mathematical expressions. In Fig.2, we represent the interference evolution as a function of the frequency difference (8). The blue curve corresponds to the interference in a realistic case. We can observe that the interference is lowered if the frequency difference δ_f of two carriers is large enough. However, we should not neglect the interference caused for high δ_f. Indeed, in the case of a high interfering number, the interference will aggregate, and can lead to errors. On the contrary, a unique user will cause a significant amount of interference only if δ_f is very small, as the filter is very selective. Thus, we can observe there are 2 main areas, whose transition occurs around 200 Hz, depending on the considered criterion. In the first area, i.e., for high δ_f, the interference level is low, and mainly concentrated around -90 dB. Contrarily, in the second area, i.e., for low δ_f, the interference level is more important (up to 0 dB when using the same frequency), and almost uniformly distributed. Nevertheless, the considered band is much larger than 200 Hz (at least 12 kHz), and thus, at this scale, the interference level can also be approximated by a constant.

Therefore, we model the interference by a rectangular function:

$$I(\delta_f) = \begin{cases} I_{max} & \text{for } \mid \delta_f \mid \leq \triangle/2, \\ I_{min} & \text{for } \mid \delta_f \mid > \triangle/2. \end{cases} \qquad (9)$$

where $\triangle$ corresponds to the width of δ_f that creates high interference level. The first line corresponds to low δ_f interferers, and the second one to high δ_f interferers.

The simplified model can be used to define the upper and the lower bound of the interference pattern. For the upper bound, the maximum level can be easily identified in Fig.2, and is set to the maximum interference power i.e., $I_{max\ up}(\delta_f = 0) = 0$ dB. On the contrary, the minimum level

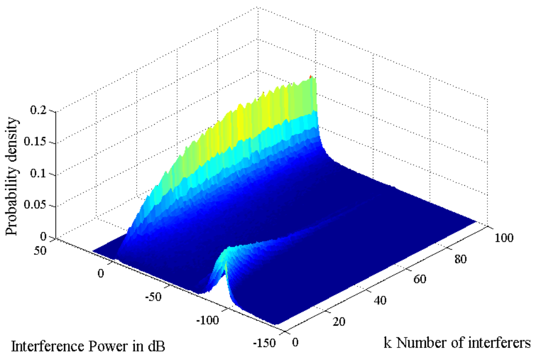


Figure 3: PDF of the aggregate interference power [dB], for $k = 100$ interferers, for BW = 12 kHz.

$I_{min\ up}$ and the width $\triangle_{up}$ can take many values, but should verify:

$$I_{min\ up} = P(\triangle_{up}) \tag{10}$$

For the lower bound, the known characteristic is the minimum level, which is set to $I_{min\ low} = -90$ dB (we neglect the lower interference values as they occur with a very low probability), whereas the other two parameters are jointly are jointly defined such as:

$$I_{max\ low} = P(\triangle_{low}) \tag{11}$$

We can also define an approximated model with unconstrained parameters $(\triangle, I_{min}, I_{max})$. We consider that $I_{min} = -90$ dB, which is the most frequent interference value, the optimal rectangular model is defined by the couple $(\triangle, I_{max})$. The bound and approximation model parameters are derived in the next section.

B. Modelization of a multi-interferers contribution

As in practice, the network will support more than 2 active users in practice, we further our study by considering more users based on R-FDMA scheme and in a realistic deployment. In this section, we aim at quantifying the cumulative interference and its influence on the system performance.

To characterize the interference statistics, we used a Monte Carlo simulation with number of repetitions: 10^4, for a network containing up to $k = 100$ interferers ($N = 101$ active nodes), deployed randomly over a continuous bandwidth of BW = 12 kHz. For the sake of simplicity, we suppose that the desired user is transmitting in the middle of the total band. Besides simplicity, this case corresponds to the worst case. Indeed, at this central frequency, the desired user will suffer from statistically more interference than any other active user. This is due to the fact that the average δ_f is smaller in this case. We have evaluated the aggregate interference power (AIP) and observe its Probability Density Function (PDF) distribution. Simulation results are presented in Fig.3.

We can verify that if the number of nodes is small, the power level of AIP remains very small and is mostly situated in the interval from -60 to -90 dB. Contrarily, when the number of node increases, the AIP gradually converges to the left, near 0 dB (which corresponds to $\delta f = 0$ for a single interferer case) and more. In fact, when the number of active users increases, the probability that at least one user chooses a frequency close to the receiver of interest is also increased. This contribution will dominate the others, and lead to a high level of interference. Finally, we can point out that the interference evolution is not trivial. Indeed, we can note 2 areas of interest (-90 dB and 0 dB) where the probability is dominant. Therefore, as shown in Fig.2 and Fig.3, the AIP cannot be approximated by a classical model, such as a Gaussian approximation for example, because, it does not take into account both main lobe for small δ_f, and side lobe for large δ_f, even for a unique interferer. But, as the interference is difficult to model exactly, we have chosen to use the rectangular model, to estimate the network AIP.

In (9), as the interference created by a unique user is supposed to take only 2 values, we distinguish 2 kinds of interferers:

– Those whose frequency shift is $\mid \delta_f \mid \leq \triangle/2$ and create interference level I_{max}. We call n_L the number of such users. The probability for an user to be in this category is $p = \frac{\triangle}{BW}$.

– The others, which create interference level I_{min}. We call $n_P = k - n_L$ the number of interferers in this case.

Thus, the total aggregate interference power I_{tot} created by k active interferes is:

$$I_{tot}(k, n_L) = n_L \cdot I_{max} + (k - n_L) \cdot I_{min} \tag{12}$$

Besides, the probability to have exactly n_L users among the k ($\forall n_L \in [0, 1, ..., k]$), that creates an interference of I_{max} is:

$$Pr(N_L = n_L) = C_k^{n_L} \cdot p^{n_L} \cdot (1-p)^{(k-n_L)} \tag{13}$$

Thus, from (4), (5) and (6), the BER and OP can be obtained with:

$$BER(k) = \sum_{n_L=0}^{n_L=k} Pr(N_L = n_L) \cdot Q\left(\sqrt{\frac{P_s}{I_{tot}(k, n_L)}}\right) \tag{14}$$

$$OP(k) = \sum_{n_L/I_{tot}(n_L)>\beta} Pr(N_L = n_L) \tag{15}$$

The (14) and (15) can be used for whichever rectangular model, in general, for the upper and lower bound, and for the approximation in particular. By using root mean square (RMS), we have evaluated the RMS_{BER} and RMS_{OP} as a function of $\triangle$ for the lower and the upper bound. Then, we have deduced consecutively the values $I_{min\ up}$ and $I_{max\ low}$ with (10) and (11). Indeed, the results using the simplified model have been compared to simulation ones (with RMS metric performed in the logarithmic scale so as to ensure a good approximation for whichever magnitude degree) to determine the best width $\triangle$ and the corresponding interference level. This study has been done for several bandwidths (BW).

As shown in Fig.4, the minimal RMS is independent of BW. For upper bound, the optimal width is obtained for $\triangle_{up} = 440Hz$ in term of both BER and OP. On the other hand, for lower bound, the optimal width in term of BER and OP will be respectively $\triangle_{low} = 100$ Hz and $\triangle_{low} = 220$ Hz. The obtained upper and lower bounds models are represented in Figure2.

We can also use these equations to empirically evaluate $(\triangle, I_{min}, I_{max})$ that are the most accurate from (9). We have evaluated the RMS_{BER} and RMS_{OP} as a function of the couple $(\triangle, I_{max})$. We have compared (with log-scale RMS metric) the BER and OP obtained with the theoretical model, and by simulation for $BW = 12$ kHz. Results obtained with a sampling precision of 1Hz and 0,005 dB are presented in Fig.5 and Fig.6. We can observe that, the width $\triangle$ has little impact on the BER accuracy, while I_{max} has little impact on

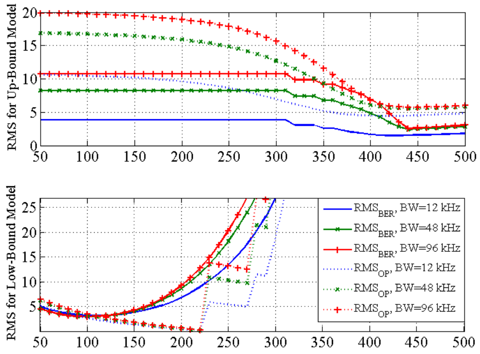


Figure 4: RMS for BER and OP vs $\triangle$, for $k = 100$, different bandwidth length.

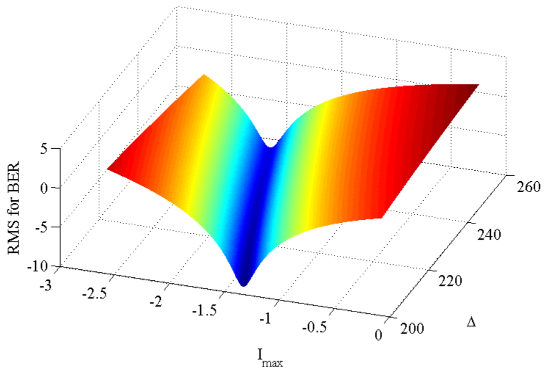


Figure 5: RMS for BER vs the couple $(\triangle, I_{max})$, for $I_{min} = -90$ dB, $k = 20$ interferers and BW=12 kHz.

the OP accuracy. Thus, regarding the OP criterion in Fig.6, we get the best approximation for $\triangle = 232Hz$. On the contrary, in Fig.5, we identified $I_{max} = -1.77$ dB as the best one for BER. Therefore, the couple ($\triangle = 232$ Hz, $I_{max} = -1.77$ dB) is considered as the optimal one (plotted in Fig.2) for both OP and BER approximation.

We validate the accuracy of our models (lower bound, upper bound and approximation) by considering a higher bandwidth, i.e., $BW = 96$ kHz. We present in Fig.7 and Fig.8 the comparison between the average BER and OP obtained by simulation, and obtained with our theoretical models. We can first verify the accuracy of the lower and upper bounds as they provide a coherent interval for the capacity. Besides, we can note that the lower bound obtained with the BER criterion is equally pertinent for the BER and OP evaluation. On the contrary, the one obtained with the OP criterion is tight for the OP, but much too loose for the BER. Finally, we can observe that the approximation model is very accurate, even for a higher bandwidth (and thus a higher supported number of users). Thus the proposed models are consistent.

IV. Estimated Capacity Network

In this section, we estimate the system capacity in terms of the maximum number of users that can be simultaneously active; while verifying the targeted BER or OP constraint. We report in Table I, Table II and Table III, the system capacity

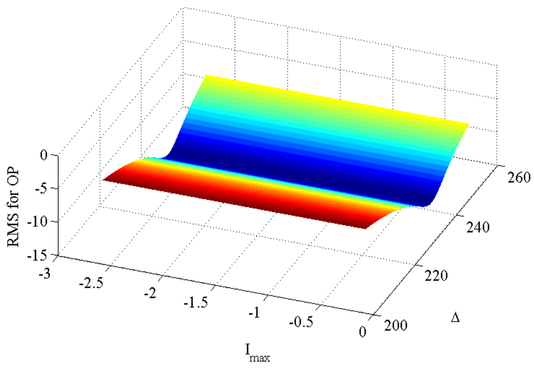


Figure 6: RMS for OP vs the couple $(\triangle, I_{max})$, for $I_{min} = -90$ dB, $k = 20$ interferers and $BW = 12$ kHz.

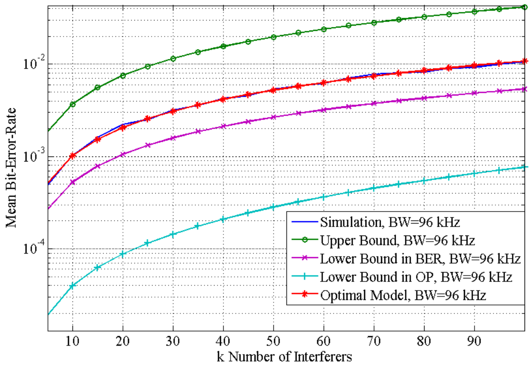


Figure 7: Mean BER as a function of k interferers, for BW = 96 kHz.

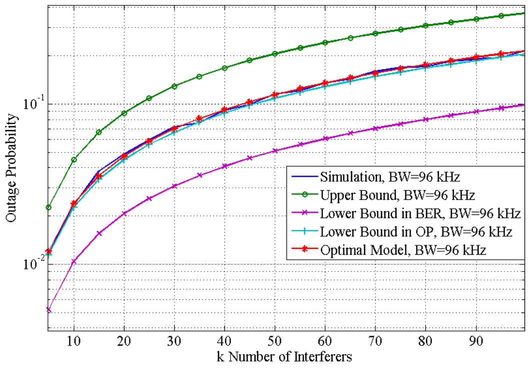


Figure 8: OP as function of k interferers, for BW = 96 kHz.

using the bounds and optimal model, and compare them with results obtained by simulation.

We can further confirm the accuracy of the bounds and optimal model. Besides, obviously, the capacity increases with the available bandwidth, and the targeted BER. However, we can note that the evolution is not linear. Indeed, e.g., when the bandwidth is increased by 8 (from 12 kHz to 96 kHz), the capacity is increased by 7.3 (from 6 to 44). Indeed, it is different to distribute N users in a B total bandwidth than

TABLE I: Maximum Transmitters Numbers For $BER = 10^{-3}$

BW	N up	N simu	N optimal	N low (BER)	N low (OP)
12 kHz	1	2	2	3	16
24 kHz	1	3	3	5	31
48 kHz	2	5	6	10	61
64 kHz	2	7	7	13	80
96 kHz	3	10	11	20	119
1 MHz	28	103	104	199	1263

TABLE II: Maximum Transmitters Numbers For $BER = 10^{-2}$

BW	N up	N simu	N optimal	N low (BER)	N low (OP)
12 kHz	4	12	13	24	63
24 kHz	7	25	24	46	124
48 kHz	14	47	48	92	244
64 kHz	18	64	63	122	323
96 kHz	27	93	94	183	479
1 MHz	157	954	976	1918	5166

TABLE III: Maximum Transmitters Numbers For $OP = 10^{-1}$

BW	N up	N simu	N optimal	N low (BER)	N low (OP)
12 kHz	3	6	6	13	6
24 kHz	6	11	11	26	12
48 kHz	12	23	23	51	23
64 kHz	16	30	30	68	31
96 kHz	23	44	45	102	46
1 MHz	124	434	455	1054	479

$N * m$ users in a $B * m$ bandwidth. Besides, with an increased number of users, some insignificant interference contributions sum up to a significant level.

Finally, we can estimate that, for a $BER = 10^{-3}$ and $BW = 96$ kHz, the network is able to serve 10 simultaneous users. Considering average transmission duration of 1 second, the system will be able to handle around 864 000 transmissions per day, which corresponds for a 50 km radius to a density of 110 nodes per km^2: i.e., 3 times the USA population density.

V. Conclusion

In this paper, we have studied a new technology based on UNB transmission, considered for IoTs networks. This technology is used jointly with R-FDMA scheme, which, to the best of our knowledge, has not been studied yet in the literature in terms of interference and capacity. To evaluate the interference impact, we have considered the BER and the OP of the system in the R-FDMA case, where the users are randomly distributed. We have studied the influence of aggregate interference power for such networks. To this aim, we have presented a rectangular model, used to derive lower bound, upper bound, and approximated model of the system. We have shown the accuracy of the models. Then, thanks to their simplicity, we have theoretically evaluated the system performance (in term of BER and OP), and the capacity of the network in terms of possible number of active users. Thus, this study is a first step in the analysis of the promising UNB networks, can be furthered by considering the case where the received powers are different among the users, to take into account the cell geometry.

Acknowledgment

The authors would like to thank Christophe Fourtet (Sigfox) his helpful advices during this study.

References

[1] L. Mainetti, L. Patrono, and A. Vilei, “Evolution of wireless sensor networks towards the Internet of Things: A survey,” 19th International Conference on Software, Telecommunications and Computer Networks (SoftCOM), Split - Hvar - Dubrovnik, Croatia, 15-17 Sept. 2011, pp. 1-6.

[2] Ye Li, Dengyu Qiao, Zhao Xu, Da Xu, Fen Miao, and Yuwei Zhang, “Energy-Model-Based Optimal Communication Systems Design for Wireless Sensor Networks,” International Journal of Distributed Sensor Networks, vol.2012, 2012.

[3] Pei Huang, Li Xiao, S. Soltani, M. W Mutka, and Ning Xi, “The Evolution of MAC Protocols in Wireless Sensor Networks: A Survey,” Communications Surveys and Tutorials, IEEE, vol.15, no.1, First Quarter 2013, pp.101-120.

[4] A. Bachir, M. Dohler, T. Watteyne, and K.K Leung, “MAC Essentials for Wireless Sensor Networks,” Communications Surveys and Tutorials, IEEE, vol.12, no.2, Second Quarter 2010, pp. 222-248.

[5] H. K. Le, D. Henriksson, and T. Abdelzaher, “A Practical Multi-channel Media Access Control Protocol for Wireless Sensor Networks,” International Conference on Information Processing in Sensor Networks, IPSN’08, St. Louis, Missouri, USA, 22-24 April 2008.

[6] Y. Wu, J.A. Stankovic, T. He, and S. Lin, “Realistic and Efficient Multi-Channel Communications in Wireless Sensor Networks,” The 27th Conference on Computer Communications. IEEE, INFOCOM 2008, Phoenix, AZ, USA, 13-18 April 2008.

[7] Q. Yu, J. Chen, Y. Fan, X. Shen, and Y. Sun, “Multi-Channel Assignment in Wireless Sensor Networks: A Game Theoretic Approach,” Proceedings IEEE, INFOCOM 2010, San Diego, CA, USA, 14-19 March 2010.

[8] M. T. Do, C. Goursaud, and J. M. Gorce, “On the Benefits of Random FDMA Schemes in Ultra Narrow Band Networks,” to be presented in International Workshop on Wireless Networks: Communication, Cooperation and Competition, WNC3’14, Hammamet, Tunisia, 12-16 May 2014.

[9] M.Z. Win, P.C. Pinto, and L.A. Shepp, “A Mathematical Theory of Network Interference and Its Applications,” Proceedings of the IEEE, vol.97, no.2, Feb. 2009, pp. 205-230.

[10] M. Haenggi, and R. K. Ganti, “Interference in Large Wireless Networks,” Foundations and Trends in Networking, vol. 3, no. 2, 2008, pp. 127-248.

[11] M. Aljuaid, and H. Yanikomeroglu, “Investigating the Gaussian Convergence of the Distribution of the Aggregate Interference Power in Large Wireless Networks,” IEEE Transactions on Vehicular Technology, vol.59, no.9, Nov. 2010, pp. 4418-4424.

[12] H. Inaltekin, and S.V. Hanly, “On the rates of convergence of the wireless multi-access interference distribution to the normal distribution,” Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), May 31 2010-June 4 2010, pp. 453-458.

[13] H. Inaltekin, “Gaussian Approximation for the Wireless Multi-access Interference Distribution and Its Application”, Journal IEEE, 2012.

[14] M. Aljuaid, and H. Yanikomeroglu, “A Cumulant-Based Characterization of the Aggregate Interference Power in Wireless Networks,” Vehicular Technology Conference (VTC 2010-Spring), 16-19 May 2010, pp. 1-5.

[15] J. Riihijarvi, and P. Mahonen, “A model based approach for estimating aggregate interference in wireless networks,” International ICST Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM), 18-20 June 2012, pp. 180-184.

[16] www.sigfox.com/en/, May 2014.

Fuzzy-based Interference Level Estimation in Cognitive Radio Networks

Minh Thao Quach, Francine Krief
LaBRI, University of Bordeaux
Talence, France
Email: {quach, krief}@labri.fr

Mohamed Aymen Chalouf
IRISA, University of Rennes
Rennes, France
Email: mohamed-aymen.chalouf@irisa.fr

Hicham Khalifé
Thales Communications & Security
Colombes, France
Email: hicham.khalife@thalesgroup.com

Abstract—**Fuzzy logic is used in various areas such as economics, train systems, smart home systems, telecommunications. Recently, fuzzy logic has attracted researchers working in cognitive radio networks (CRNs). In this paper, we introduce a method that uses fuzzy logic to combine observed factors of the wireless environment (e.g., area overlapping and primary receivers density) to estimate interference level to primary receivers. The computed results reflect the precise impact that may be induced when a cognitive radio communication is operating nearby. This impact envisages the effects of CRNs over primary receivers. It can also be used as a routing metric that helps to choose the route with minimal impact - lowest interference level - to primary receivers.**

Keywords–Cognitive radio; fuzzy logic; routing metric; interference avoidance.

I. INTRODUCTION

In a Cognitive Radio Network (CRN), a cognitive radio node (CR) makes decisions based on its own observed information even though these knowledge may be incomplete. Fuzzy logic, however, can yield useful outputs with incomplete approximate and vague information (e.g., low or high interference, sufficient or not sufficient available radio resources). Furthermore, fuzzy logic does not require too complicated computation since the calculation is mostly based on If-then-else rules. Hence, we can use fuzzy logic in real-time cognitive radio applications for which the response time is crucial to the system performance [1]. Due to its simplicity, flexibility, and if-then-else rules composition, fuzzy logic processing time is minor.

Fuzzy logic introduces a logic theory that was developed to generalise 'true' and 'false' values to any value between 0 and 1 [2]. It also presents the approximate knowledge which may be difficult to express by conventional crisp method (i.e., bivalent set theory).

A fuzzy logic system with two inputs and one output is described in Figure 1. The fuzzy sets are sets of unsharp boundaries objects in which the membership is a matter of degree (in range of 0 to 1). For instance, a fuzzy set of $weekend$ may contain half of Friday, Saturday and Sunday and a set of $weekdays$ may contain from Monday to first half of Friday. So, Friday can be existing in both sets with distinctive degrees. To identify the degree of these variables, a membership function is used to reason the related information. The membership function assigns a value in the interval $[0, 1]$ to a fuzzy variable and denotes as $\mu(weekend(day))$, where $weekend$ is a fuzzy set, and day is a fuzzy variable.

Input crisp values are fuzzified to produce appropriate linguistic values according to defined membership functions. Then, the inference engine will extract the associated outputs based on the defined rules. These outputs are fuzzified based on output membership functions. Finally, fuzzified outputs are aggregated into a single crisp value by the defuzzifier.

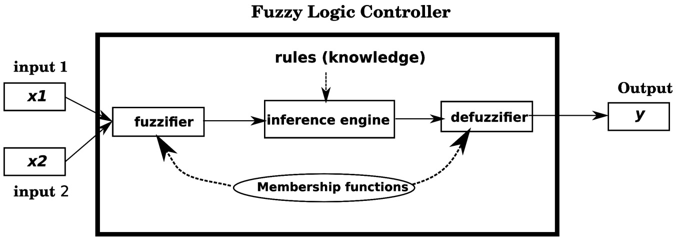


Figure 1: General fuzzy logic system

In the previous articles [3] [4], we showed that reception overlapping associated with the interference could impact the primary radio (PR) receivers. However, we also noticed the case where node density also contributed to the impact. It is worth mentioning that we only consider non-zero overlapping situation since zero overlapping does not impact to the primary network.

The output can be used to investigate how the routing layer reacts and makes the right decisions to maximise spectrum resources while avoiding interference to the primary receivers. For instance, a CR can operate within an area having high overlap size but low operating primary receivers. We apply fuzzy logic to determine the overlap size and the probability of operating primary receivers (e.g., low or high). We also introduce in details the methodology of Mamdani inference system so that the audiences can easily follow the proposed solution.

The rest of the paper is organised as follows. Section II presents some recent work related to fuzzy logic applications. Our basic implementation is then discussed in Section III. The advance implementations in Cognitive Radio Networks (CRNs) context is presented in Section IV. We conclude the work with some future directions in Section V.

II. RELATED WORK

Many fuzzy logic based solutions have been proposed. For instance, fuzzy-decision based routing was introduced in [5] for MANET. It was developed on top of the classical Dynamic Source Routing (DSR) protocol in order to achieve the fairness of all the routing input metrics and prioritize the services differentiated packet routing, i.e., to route packets based on QoS priority. Wong et al. [5] proposed a routing protocol based on fuzzy decision engine that supports service differentiation and quality of service (e.g., routing protocol with service engineering or service-based routing in MANET).

Rea et al. [6] used fuzzy logic to instruct route caching during path exploration process to ensure that only the quality routes are cached in DSR. The solution also uses hop count as one of the metrics similarly to [7]. Furthermore, it uses link strength and energy available at a link vertex as fuzzy inputs. The outcome is whether a path in a route request is cached or not and continue with a route request rebroadcast. Though the solution at that stage was still not completely dealing with changes of the wireless environment, applying fuzzy theory in ad-hoc routing is convincing.

Chiang and Wang [8] used fuzzy logic to optimize routing path in a distributed manner. The objective of the proposed protocol is to minimize resource energy consumption in order to lengthen the lifetime of the sensor network. Since fuzzy logic is a system based on a conditional statements rule, resources consumption was reduced as expected in this proposal.

Santhi et al. [9] applied fuzzy logic to combine different QoS criteria to produce a routing metric. Ad-hoc mobile node uses this metric to predict and choose the most stable but least cost path to reach the destination. Similar approach can be used in CRNs but we have to consider the surrounding‘s changes that affect the routing decision as well as routing performance, such as the resource availability and operation interference on the legacy primary systems.

In CRN design development, Baldo et al. [10] suggested to use fuzzy logic in controlling transmitting power of the CR devices while it co-exists with PR devices. Le et al. [11] proposed a design of network accessing scheme based on fuzzy logic. Fuzzy logic was used to combine multiple feedbacks of a device on network performance such as delay, throughput and reliability. The output of the network selection outperformed conventional scheme on choosing an access based on a single parameter.

Masri et al. [12] proposed a strategy that used fuzzy logic to compose multiple independent environment parameters for multihop routing in CRNs. They accounted for instantaneous variations of the environment and proved that channel selection must be part of routing decision jointly with MAC layer support. In this work, we aim to extend the work described in [12] by adding impact factors that are derived from the environment observation. The solution proposes a metric that could guarantee the minimal impact to the primary system when it coexists with the secondary system.

III. Basic implementation and results

We argue that when reception area of a CR emitter and reception area of a PR emitter overlap, it produces unavoidable effects on the primary system, especially to the PR receivers. This observation was mentioned in [3]. However, we also proved that not only the overlap size but also the number of existing primary receivers cause the impact [4]. In this work, our approach takes into account the overlap ratio and node density probability as two main factors. The ratio of the overlap size over the overall size of the PR emitter's disk, named overlap ratio while node density probability is the probability of possible node density within this PR emitter's disk.

We hence choose overlap ratio and node density probability to be the fuzzy inputs of our fuzzy inference system. Each of these two variables is composed of two fuzzy sets, i.e., *Low* and *High*. The fuzzy output variable is the interference level that is also a fuzzy set containing two fuzzy variables *Low* and *High*. A proper implication would be applied for each rule listed in the rules table. Result from implication rule is then aggregated and defuzzified to obtain the final result. This is the degree of impact on the primary system. A CR can consider this degree before using a frequency range when overlap happens.

A. Overlap Ratio Fuzzy Sets

Ratio of an overlap area to reception zone of an emitter is taken as the fuzzy input variable. To make it simple and easy to understand, we first define two simple fuzzy sets *Low* and *High* that represent the overlap ratio state. *Low* set contains all values that indicate the low overlap ratio. For instance, overlap ratio is considered low when it is less than 50%, otherwise, it is considered high. Note that, specific low and high boundaries are not precisely defined. We can have different definitions of *low* and *high*. We are declaring the simplest possibility in this context. It is said to be 100% low when the ratio is exactly from 0 to 20%. From 20% to further (e.g., 50%), the possibility of being low hence decreases while the possibility of being high increases. We describe the membership functions of these sets in trapezoidal or triangular-shape. Overlap ratio membership functions are described in equations (1) and (2).

$$\mu_{Low}(x) = \begin{cases} 1 & 0 \leq x \leq 20\% \\ \frac{50\% - x}{30\%} & 20\% < x \leq 50\% \\ 0 & x \geq 50\% \end{cases} \tag{1}$$

$$\mu_{High}(x) = \begin{cases} \frac{x - 25\%}{50\%} & 25\% \leq x \leq 75\% \\ 1 & x \geq 75\% \end{cases} \tag{2}$$

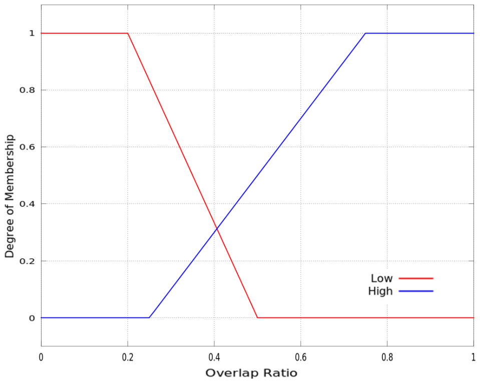


Figure 2: Membership function of Overlap Ratio.

To interpret the output of antecedents (i.e., the overlap ratio), we use Mamdani Min Implication rules [2] to extract the final result for the overlap ratio fuzzy set. For instance, at intersection part of two functions, an *or* operator is used to connect two sets, the maximum of two membership functions is evaluated for the antecedent part of the fuzzy rules.

$$\begin{aligned}\mu_{OverlapRatio}(x) &= \mu_{Low}(x) \vee \mu_{High}(x)\\ &= max[\mu_{Low}(x), \mu_{High}(x)] \qquad (3)\end{aligned}$$

B. Node Density Fuzzy Sets

Similarly to Overlap Ratio sets, we also define two simple fuzzy sets *Low* and *High* in order to reflect how PR receivers are scattered within an area. The characteristics of the mobile receivers are discrete, independent and randomly distributed. Therefore, the distribution of the nodes in this context is assumed to follow Poisson distribution. The expected density value X yields from Grey Model [4]. Since the node density appearance is a mutual independent event occurring at a known and constant rate r per unit (of time or space) are observed through a certain window (a unit of time or space), it follows the principle of Poisson distribution.

From the historical data (i.e., the input series for Grey Model), we can compute the possible average density m that represents the estimated rate λ. The probability the area has at most X receivers within an area unit is the Poisson accumulate density function of $P(X \leq x)$ illustrated in (III-B), $P(X \leq x) = \frac{e^{-\lambda}\sum_{i=0}^{x}\lambda^{i}}{i!}$

Higher probability of predicted density is, higher chance the receivers get impact. We consider that the low value of $P(x)$ belongs to fuzzy set *Low* and the other belongs to fuzzy set *High*. The membership functions of Node Density are also presented in a trapezoid-shape similarly to Overlap Ratio fuzzy sets

C. Fuzzy Process for Overlap and Node Density sets

Since Overlap ratio and Node Density are two independent entities with different properties and characteristics, we combine these two sets using a rules table. Output of the combination represents the interference level (e.g., *Low* or *High* level of interference) to the primary system under specific overlap degree $\mu_{OverlapRatio}(ratio)$ and primary receiver density degree - defined as $\mu_{Density}(P_x)$. The interference level is used to foresee how much impact primary receiver would be tolerated, density of these nodes are hence prioritized in this rules table (Table I).

TABLE I. INTERFERENCE LEVEL RULES TABLE

Overlap Degree	Density Degree	Interference Level
Low	Low	Low
High	Low	Low
Low	High	High
High	High	High

The rules table is expressed in the If-then construct. For instance, if both the overlap ratio and the density are *Low*, interference level is *Low*. However, interference level is *Low* when overlap ratio is *High* and the density is *How*. This explains the case where we have big overlap and low receivers operating in the emitter's reception zone. The statement if-part of the rule is called *antecedent* or premise, while the then-part of the rule is called *consequence* or conclusion. In this context, the premises are the overlap ratio and the density degree, while the consequence is the interference level. The consequence is also a fuzzy set. We define the fuzzy sets of interference level in Figure 3.

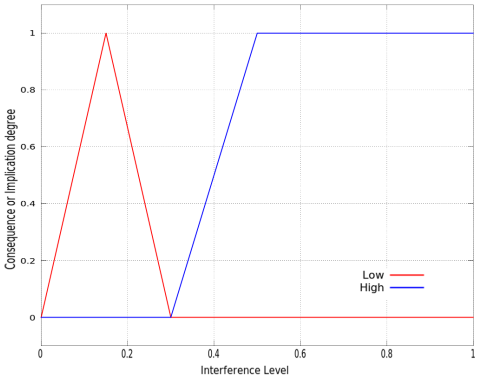


Figure 3: Membership function of the output Interference Degree.

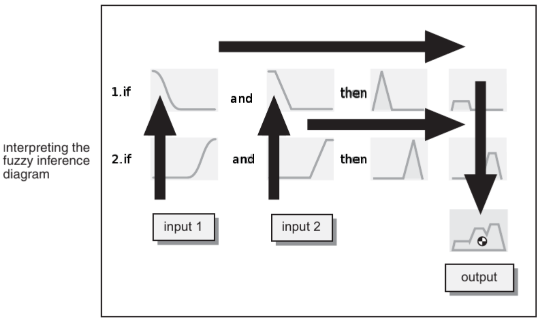


Figure 4: Fuzzy Inference Mapping Diagram.

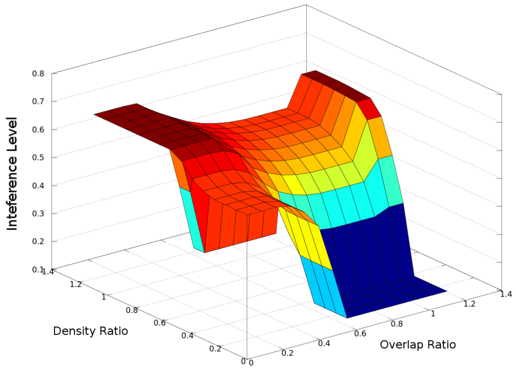


Figure 5: Interference Level Output as function of Overlap and Node Density - Rule set of 4.

In general, the inputs to these rules are the current values of overlap and density degrees and the output would be the entire fuzzy set of interference level (e.g., Low or $High$ set). This set is then defuzzified that assigns a single value to indicate the interference ratio. The mapping is done from-left-to-right flows as shown in Figure 4 [13]. The graphical view of the interference level according to the overlap ratio and the node density is presented in Figure 5.

TABLE II. OVERLAP DEGREE FUZZIFICATION OUTPUT

Index	Overlap Ratio	$\mu(x)$ Low	$\mu(x)$ High	$\mu(x)$ Overlap Ratio	Overlap Degree
1	0.06718	1	0	1	Low
2	0.25534	0.81552	0.01069	0.81552	Low
3	0.42753	0.24155	0.35507	0.35507	High
4	0.65681	0	0.81362	0.81362	High

Table II represents the numerical data after fuzzification and defuzzification process of Overlap Ratio fuzzy sets. $Ratio$ is the overlap ratio of overlap region to the emitter reception zone. This is the fuzzy input of fuzzification process to map this value to the linguistic variables *Low* and *High*. Fuzzy logic controller (FLC) uses the membership functions defined in Figure 2 to fuzzify the input ratio. $\mu(x)Low$ and $\mu(x)High$ present the fuzzified values obtained from equations (1) and (2) respectively. For example, at index 2 input ratio that equals 0.25534 is interpreted as 81.56% Low and 1.07% High according to the membership functions defined in Figure 2. These outputs are evaluated as 81.56% low overlap (e.g., columns $\mu(x)$ Overlap Ratio and Overlap Degree) by Mamdani Min Implication in equation (3).

TABLE III. DENSITY DEGREE FUZZIFICATION OUTPUT

Index	Poisson Distribution of Node Density	$\mu(x)$ Low	$\mu(x)$ High	$\mu(x)$ Density	Density Degree
1	0.26119	0.79601	0.0224	0.79601	Low
2	0.89204	0	1	1	High
3	0.94045	0	1	1	High
4	0.77162	0	1	1	High

The same processes are done and presented in Table III. Density degree and overlap degree are composed by applying Larsen implication rule with the rule explained in Table I. *Low* overlap ratio combined with *low* density probability would result a *low* interference level. Defuzzified value of this result is produced by multiplying the crisp values of overlap ratio and density probability. Table IV shows the final output result after aggregating the two fuzzy sets Overlap Degree and Density Degree. We can see that if overlap is *low* and density degree is *low*, interference level is hence *low* (first row of the table in Table IV). Crisp value column represents defuzzified output of overlap and density degree sets. Interference level is at 79.6% *low* when overlap degree is 100% *low* and density degree is 79.6% *low* for instance at index 1. The corresponding outputs of overlap and density degree antecedents are presented in Table II and Table III.

TABLE IV. INTERFERENCE LEVEL FUZZIFICATION OUTPUT

Index	Overlap Degree	Density Degree	Interference Level	Crisp value
1	Low	Low	Low	0.79601
2	Low	High	High	0.81552
3	High	High	High	0.35507
4	High	High	High	0.81362

With this simple approach, we can see that interference level depends on the predicted density degree. However, it provides the glimpse of considering overlap and density for better protecting primary receivers in CRNs. Moreover, considering binary variables as Low and $High$ is not sufficient enough to evaluate how interference level is good enough to make a judgement when it comes to path selection. Clearly, the rule table as well as the outcome of the defuzzification process does not reflect any impact of the overlap size. A middle level of impact may happen due to low overlap even with high density receivers. An extension of this approach is hence introduced briefly below. This enhances overlap and density fuzzy sets, as well as refines rules table, accordingly.

IV. Advance Interference Level Implementation

A. Extended Overlap Ratio Fuzzy Sets

As explained, we need more elaborate definition for the input of FLC that covers all possible cases of the overlap and density effects. We redesigned the input fuzzy sets that now contain four linguistics variables $O_{fuzzyset} = \{low, medium, high, veryhigh\}$ in which, for example, the overlap ratio:

- $low : 0 < \frac{A_{overlap}}{A_P} \leq \frac{3}{10}$
- $medium : \frac{2}{10} < \frac{A_{overlap}}{A_P} \leq \frac{5}{10}$
- $high : \frac{4}{10} < \frac{A_{overlap}}{A_P} \leq \frac{7}{10}$
- $veryhigh : \frac{A_{overlap}}{A_P} > \frac{6}{10}$

We consider that the overlap ratio is low when the ratio is from 0 to 10%. The membership function of *low* overlap ratio reflects the degree of low overlapping which follows the idea of the possibility that overlap is lower and lower after 10% of overlap. The possibility of *low* overlap decreases when the ratio increases. Other possibilities could be *medium* or *high* overlap after a specific boundary. For instance, the input value of an overlap ratio is at 25%, the probability of it to *low* is 25%. Moreover, the probability of it to *medium* is 33%. We can conclude that the input may be possibly at *medium* overlap degree.

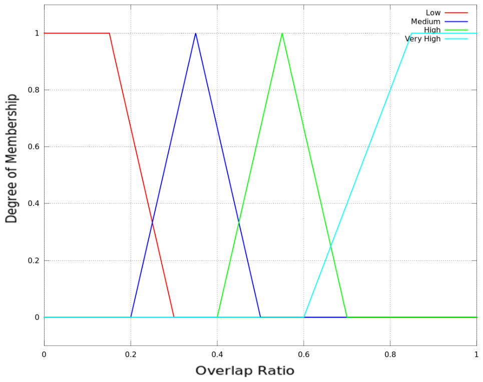


Figure 6: Enhanced Overlap Ratio membership function.

B. Extended Node Density Fuzzy Sets

Applying the same principle in IV-A, we define four linguistic variables corresponding to four fuzzy sets *Low*,

ISBN: 978-1-61208-360-5

Medium, *High* and *Very High*. Assume that density of the receivers is defined as $d_r = \frac{N}{A_P}$.

where N, the total number of receivers within the reception zone of an emitter, A_P the total area size of the zone. Therefore, within a specific area A_0, the average number of receivers is $n_{avg} = d_r * A_0$.

As a node existence is independent from the others within an area, the probability of x nodes which exist within A_0 follows Poisson process with mean $\lambda = n_{avg}$ is yield by (4) following the Poisson density function.

$$f(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \tag{4}$$

The probability the area has more than the average number of receivers within A_0 is the Poisson accumulate density function of $P(X \geq x)$ with $x \geq \lambda$ becomes $P(X \geq x) = \frac{e^{-\lambda}(e\lambda)^x}{x^x}$

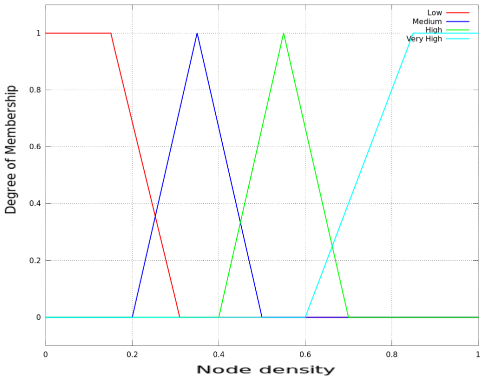


Figure 7: Enhanced Node Density membership function.

Practically, we can have a prediction system that generates the node density. However, in the context of this paper, the value is generated via a Poisson process. This value is converted into an appropriate linguistic value via the fuzzy logic controller. The membership function of this fuzzy set follows the definition from the overlap membership function above as described in Figure 7.

C. Interference Level Rules and Outputs

As the inputs are refined, the output of this enhanced inference system is also refined. A proposed rule table for these fuzzy sets is defined in table V, note that the rules subject to feasibly change depending on realistic observation later. As we could see, with the simple approach in section III, the rules illustrate only two states of the interference level, *low* or *high*. However, we introduced two more variables of reach input fuzzy sets (*medium* and *veryhigh*), the rules should also reflect the changes associated with these variables.

Practically, the level of interference can be at a reasonable degree such as medium. Based on application needs, the level could be adapted accordingly.

For instance, the above rules infer the followings.

- If (Overlap-Ratio is *Low*) or (Density-Ratio is *Low*), then (Interference Level is *Low*) (1)

TABLE V. ENHANCE INTERFERENCE LEVEL RULES TABLE

Index	Overlap Ratio	Density	Interference Level
1	Low	Low	Low
2	Low	Medium	rather Medium
3	Low	High	somewhat High
4	Low	Very High	High
5	Medium	Low	somewhat Low
6	Medium	Medium	Medium
7	Medium	High	High
8	Medium	Very High	Very High
9	High	Low	Medium
10	High	medium	somewhat High
11	High	High	High
12	High	Very High	Very High
13	Very High	Low	Medium
14	Very High	Medium	very High
15	Very High	High	extremely High
16	Very High	Very High	extremely very High

- If (Overlap-Ratio is *Low*) and (Density-Ratio is *Medium*), then (Interference Level isn't *Low* or rather *medium*) (0.5000)
- If (Overlap-Ratio is *High*) or (Density-Ratio is *Medium*), then (Interference Level is somewhat *High*) (1)
- If (Overlap-Ratio is *High*) or (Density-Ratio is *High*), then (Interference Level is *High*) (1)
- If (Overlap-Ratio is Very *High*) and (Density-Ratio is *High*), then (Interference Level is extremely *High*) (1)
- If (Overlap-Ratio is Very *High*) or (Density-Ratio is Very *High*), then (Interference Level is extremely Very *High*) (1)

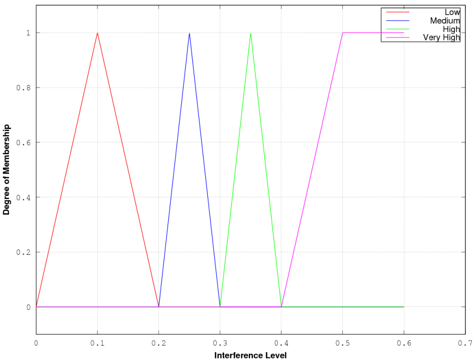


Figure 8: Enhanced Interference Level membership function.

The fuzzy inference engine combines the rules to obtain the aggregated fuzzy output. The output is the fuzzy set of the interference level that is defined in Figure 8. Fuzzy controller has to defuzzify these outputs into crisp values using centroid method to make the final decisions. Figure 9 shows the system output as a function of 2 variables, overlap ratio and node density, with the rules set of 16 conditional statements.

Table VI and Table VII show the fuzzified data of the inputs based on their defined membership functions. We can observe that the crisp values are converted into linguistic values thanks to the membership functions in Figure 6 and Figure 7. Table VIII shows the output of the inference system. The aggregated values are processed according to the Interference

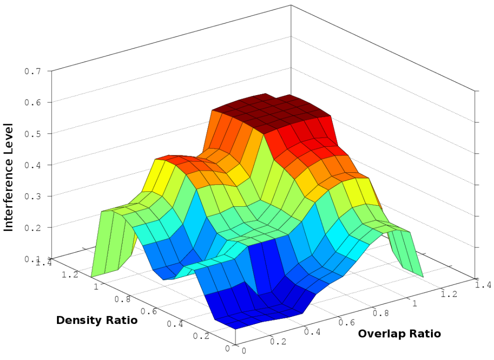


Figure 9: Interference Level Output as function of Overlap and Node Density - Rule set of 16.

Membership function in Figure 8. The rules are applied correspondingly in defuzzification process to produce the final crisp value of the Interference level.

TABLE VI. OVERLAP DEGREE FUZZIFICATION OUTPUT

Index	overlap ratio	$\mu(x)$ Low	$\mu(x)$ Medium	$\mu(x)$High	$\mu(x)$ Very High
1	0.25596	0.29362	0.37305	0	0
2	0.57829	0	0	0.81143	0
3	0.44402	0	0.37322	0.29345	0
4	0.66566	0	0	0.22894	0.26264
5	0.53420	0	0	0.89467	0
6	0.65453	0	0	0.30314	0.21811

TABLE VII. DENSITY DEGREE FUZZIFICATION OUTPUT

Index	Density probability	$\mu(x)$ Low	$\mu(x)$ Medium	$\mu(x)$ High	$\mu(x)$ Very High
1	0.08049	1	0	0	0
2	0.19908	0.67281	0	0	0
3	0.66764	0	0	0.21574	0.27056
4	0.81842	0	0	0	0.87370
5	0.28182	0.12118	0.54548	0	0
6	0.57480	0	0	0.83466	0

Precisely, FLC maps the fuzzified outputs (a.k.a. the output of each linguistic variable of overlap ratio and node density probability) of the inputs to infer the associated consequences. For instance, the inference engine decomposes an input of overlap ratio of 0.25596 into $\mu(low)$ at 0.29362 and input of density probability of 0.08049 into $\mu(low)$ at 1. This composition matches the first rule in Table V - *If (Overlap-Ratio is Low) or (Density-Ratio is Low), then (Interference Level is Low).*

Depending on the method that we defined at the beginning, the consequence of these antecedents is calculated and mapped to the membership functions of the interference level fuzzy set. With this current example, we opt to use Min implication method to evaluate the outcome, and compute the interference output is at 0.14236 for this rule. However, this is not yet the final outcome since after matching all the possible fuzzified inputs with the rule knowledge, all the outcomes are decomposed/defuzzified to produce a single crisp value (refers to the diagram in Figure 4). This will be the final output of the whole process.

V. Conclusion

In this paper, we provide an approach to estimate the interference level based on fuzzy logic. Overlap radio and node density are two critical inputs of the fuzzy system. Convincing numerical results confirm the feasibility of using fuzzy logic in Cognitive Radio Networks for estimating interference. The decision making process can leverage this information when a CR selects a possible accessing channel that minimizes the impact on the primary system. Moreover, the estimation can also be used in extracting a routing metric that considers a path with minimal interference level. In cross-layer design, this approach can also be integrated with other factors such as transmitting power and application requirements for engineering the traffic flow accordingly.

TABLE VIII. INTERFERENCE LEVEL FUZZIFICATION OUTPUT

Index	Overlap input	Density input	Interference Level	Crisp Value
1	0.25596	0.08049	Low	0.14236
2	0.57829	0.19908	Medium	0.21464
3	0.44401	0.66764	High	0.39495
4	0.66566	0.81842	Very High	0.47103
5	0.53420	0.28182	Medium	0.27273
6	0.65453	0.57480	High	0.37930

Acknowledgment

This work is supported by LICoRNe project, funded in part by the National Agency for Research in France - ANR (Agence Nationale de la Recherche).

References

[1] E. Hossain, D. Niyato, and Z. Han, Dynamic spectrum access and management in cognitive radio networks. Cambridge University Press Cambridge, 2009.

[2] T. J. Ross, Fuzzy Logic with Engineering Applications. John Wiley & Sons, Ltd., 2010.

[3] M. T. Quach and H. Khalife, "The impact of overlap regions in cognitive radio networks," in Wireless Days (WD), 2012 IFIP, 2012, pp. 1–3.

[4] M. T. Quach, D. Ouattara, F. Krief, H. Khalife, and M. A. Chalouf, "Overlap regions and grey model-based approach for interference avoidance in cognitive radio networks," in Ubiquitous and Future Networks (ICUFN), 2013 Fifth International Conference on, 2013, pp. 642–647.

[5] Y. F. Wong and W. Wong, "A fuzzy-decision-based routing protocol for mobile ad hoc networks," in Networks, 2002. ICON 2002. 10th IEEE International Conference on, 2002, pp. 317–322.

[6] S. Rea and D. Pesch, "Multi-metric routing decisions for ad hoc networks using fuzzy logic," in Wireless Communication Systems, 2004, 1st International Symposium on, 2004, pp. 403–407.

[7] G. Alandjani and E. Johnson, "Fuzzy routing in ad hoc networks," in Performance, Computing, and Communications Conference, 2003. Conference Proceedings of the 2003 IEEE International, 2003, pp. 525–530.

[8] S.-Y. Chiang and J.-L. Wang, "Routing analysis using fuzzy logic systems in wireless sensor networks," in Knowledge-Based Intelligent Information and Engineering Systems. Springer, 2008, pp. 966–973.

[9] G. Santhi and A. Nachiappan, "Fuzzy-cost based multicast qos routing with mobility prediction in manets," in Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on, 2012, pp. 556–562.

[10] N. Baldo and M. Zorzi, "Cognitive network access using fuzzy decision making," Wireless Communications, IEEE Transactions on, vol. 8, no. 7, 2009, pp. 3523–3535.

[11] H.-S. Le and Q. Liang, "An efficient power control scheme for cognitive radios," in Wireless Communications and Networking Conference, 2007.WCNC 2007. IEEE, March 2007, pp. 2559–2563.

[12] A. E. Masri, N. Malouch, and H. Khalife, "A fuzzy-based routing strategy for multihop cognitive radio networks." IJCNIS, vol. 3, no. 1, 2011, pp. 74–82.

[13] G. of Authors from the Mathworks., Fuzzy Logic Toolbox tm - User's Guide, R2013b, Ed. The Mathworks, Inc., 2013.

Performance of the LAD Spectrum Sensing Method in Measured Noise at Frequency Ranges between 10 MHz and 39 GHz

Johanna Vartiainen and Risto Vuohtoniemi
Centre for Wireless Communications, University of Oulu
Oulu, Finland
Email: firstname.lastname@ee.oulu.fi

***Abstract*—Spectrum sensing is a low-complex and interesting way to find unused white spaces for secondary users transmission in cognitive radio. Because radio frequencies are strategic resource, their reallocation is required to ensure enough capacity for future communication devices. In addition to commonly used frequencies, millimetric waves have been proposed to be used for communication to fulfill upcoming needs. Because of broad operating area, adaptive spectrum sensing methods are needed to manage in different noise environments. For that reason, noise measurements were performed at several frequency areas between 10 MHz and 39 GHz. The goal was to study the noise properties in different frequency areas. Statistical properties of measured noise areas were analyzed and compared also with theoretically generated noise. The results show that histograms, PSDs and CDFs are almost equal. However, it was noticed that there is a huge difference between the noise levels, so sensing method that adaptively sets the detection threshold is required. The localization algorithm based on the double-thresholding (LAD) method was used as a blind and adaptive sensing method. The LAD method is based on the assumption that the noise is Gaussian. The probability of detection and false alarm were studied. It was shown that the LAD method operates well in all studied frequency areas.**

Keywords–noise measurement, cognitive radio, spectrum sensing, millimetric waves.

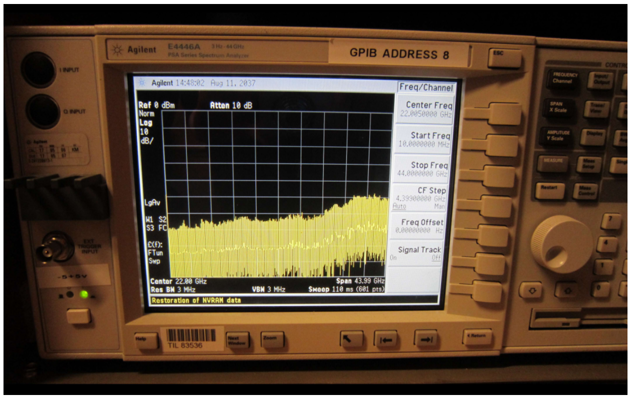


Figure 1: Agilent E4446A spectrum analyzer.

I. Introduction

In the modern information society, radio spectrum is a basic and essential element. The demand of more frequencies because of developing communications requires effective and improved resource allocation as well as novel way of thinking. More capacity is required, so existing frequency bands should be used more efficiently. One possibility is to utilize cognitive radio systems (CRS) [1][2][3][4]. In CRS, secondary (S) users can temporarily use unused white spaces aka holes in time/frequency domain where primary (P) users are non-active. In addition, more band is required. Future solution for wider frequency band demand is to use higher frequencies like millimetric waves, i.e., bands from 10 GHz-70 GHz. However, those bands require a lot of investigation and possible whole new technologies.

In CRS, unused white spaces in the spectra can be found using spectrum sensing. Even though there are also other techniques as databases, sensing is very attractive because it can be done blindly and easily. Even though the Federal Communication Commission (FCC) has decided that sensing is not required defining TV white spaces [5], sensing has a future, for example, in other frequency areas and in wireless local area network (WLAN)-type solutions when the distances between the transmitter and receiver are short and transmit power are small. In addition, public safety applications when infrastructure is down and there is no connection to databases, sensing may be needed.

Spectrum use measurements are very important to characterize white spaces for CRS. In many cases, these spectrum use measurement campaigns have used conventional spectrum analyzer as, for example, in [6][7]. The classification into signal and noise has been done with a non-adaptive single (power) threshold. The performance of this signal classification is decreased when the noise spectral density is not flat inside investigated frequency range. In some cases, radio frequency (RF) sensor has been used [8]. In the future, possible frequencies for CRS operation cover from megahertz to tens of gigahertz. The problem is that noise properties vary in different frequency areas. Thus, a critical issue to deal with that wide operating area is to adapt parameters in the different environments. Inside of broad operating frequency range there is a quite large variation in internal noise level of a conventional spectrum analyzer and an RF sensor. For example, in some sensor applications the aim is to get consistent group delay, which is a requirement for good Time Difference of Arrival (TDOA). The downside is that the noise floor inside of a receiver is not flat inside wide operating frequency range. In Fig. 1, the internal noise level of spectrum analyzer as a function of frequency is shown. It can be seen that there is over 20 dB difference between the internal noise levels at low frequency compared to noise level at higher frequencies. It is also seen that there is some noise level fluctuation also inside of much narrower bandwidth. This noise fluctuation decreases the performance of sensing if this fluctuation is not taken into account.

In this paper, several 100 MHz noise measurements at broad spectrum range from 10 MHz to 39 GHz were performed and the characteristics of measured noise at different frequency areas were analyzed. Theoretically generated noise following Gaussian distribution was used as a point of comparison. Histogram, power spectral density (PSD) and cumulative distribution function (CDF) were studied. In addition, the blind and adaptive spectrum sensing method called the localization algorithm based on double-thresholding (LAD) method [9] was used to find signals present. The LAD method is based on the assumption that the noise is Gaussian, and it determines the noise level. Here, we studied how the LAD method is able to operate in all measured noise areas. Probability of detection and false alarm were studied for measured noise areas as well as for theoretically generated noise.

This paper is organized as follows. In Section II, the LAD method is presented. Section III presents measurement setup. Noise measurement results are presented in Section IV and conclusions are drawn in Section V.

II. The LAD method

The LAD method [9] uses two forward consecutive mean excision (FCME) thresholds [10]. The adaptive FCME algorithm calculates the detection thresholds using pre-determined threshold parameter that is calculated based on the distribution of the noise. It is assumed that the noise is white Gaussian process with the one-sided power spectral density N_0. Thus, the threshold parameter T_{CME} can be found solving [11] [12]

$$P_{FA,DES} = e^{-(T_{CME}M)} \sum_{i=0}^{M-1} \frac{1}{i!}(T_{CME}M)^i, \tag{1}$$

where $P_{FA,DES}$ is the desired false alarm probability like in constant false alarm rate (CFAR) systems at M element antenna array. Note that it does not depend on the variance [11]. Here, $M = 1$, so (1) reduces to $P_{FA,DES} = e^{-(T_{CME})}$, from which we get that

$$T_{CME} = -ln(P_{FA,DES}). \tag{2}$$

Let us assume that there are N samples x_i arranged into an ascending order from smallest to largest so that $x_1 < x_2 < \ldots < x_N$. Signal samples are found iteratively searching the smallest k, $k \geq round(0.1N)$ so that [12]

$$y_{k+1} \geq T_{CME} \sum_{i=1}^{k} y_i = T, \tag{3}$$

where $y_i = |x_i|^2$ (=energy). In the first iteration, $k = round(0.1N)$ so that $\sum_{i=1}^{k} y_i$ includes 10% of the smallest samples (so called initial set assumed to consist only noise samples). Now, energy of the noise samples y_i follow the central chi-square distribution with $2M = 2$ degrees of freedom. In general, the probability density function for the chi-squared distribution with r degrees of freedom is [11]

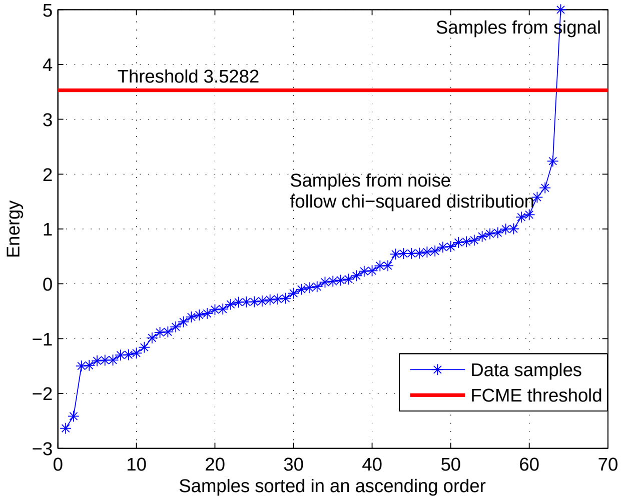


Figure 2: An example of the FCME algorithm. Impulsive signal, noise and FCME threshold. $N = 64$ samples.

$$P_r(x) = \frac{x^{\frac{r}{2}-1} e^{\frac{-x}{2}}}{\Gamma(\frac{1}{2}r) 2^{\frac{r}{2}}}, \tag{4}$$

where Γ is a gamma function. If (3) holds, y_{k+1} and values above that are decided to be from signal(s) and y_k and values below that are from the noise, i.e., the FCME algorithm estimates the noise level (Fig. 2). Thus, the samples have been divided into two sets using the threshold T:

$$\begin{aligned} y_1, \ldots, y_k &\rightarrow \text{ noise samples} \\ y_{k+1}, \ldots, y_N &\rightarrow \text{ signal samples} \end{aligned}$$

Because of the initial set assumption, the FCME algorithm assumes that at least 10% of the samples are from the noise, so at most 90% of the samples can be from the signal(s). However, the less signal samples the better the FCME algorithm operates [13].

The LAD method [9][13] calculates two FCME thresholds using two different threshold parameters T_{CME}. After that, all the adjacent samples above the *lower* threshold are grouped together to form a group G_i, $i = 1, \ldots, h$, where $h < N$. If at least one sample of each group G_i exceeds also the *upper* threshold, the group is accepted to be from the signal. If not, the group is from the noise and rejected. The number of accepted groups is $l \leq h$. The computational complexity of the FCME and LAD methods is of the order of $N \log_2 N$ [14]. An example of the LAD method is presented at Fig. 3.

III. Measurement Setup

In the noise measurements, we used high-performance spectrum analyzer (Agilent E4446A) [15]. The input signal was downconverted and digitized with 14 bit analog to digital

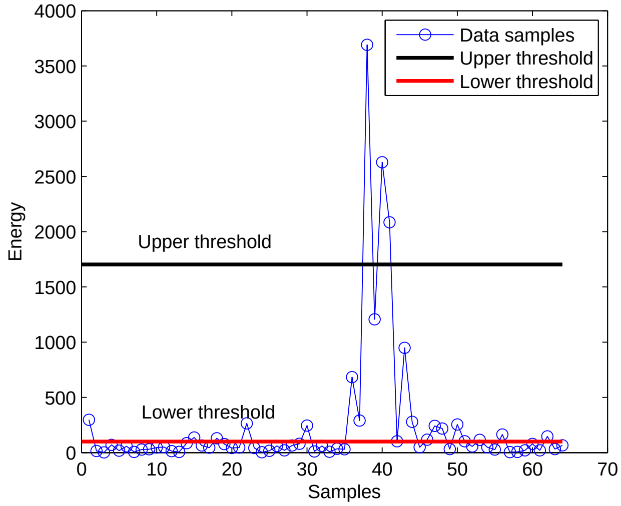


Figure 3: An example of the LAD method. BPSK signal with SNR=5 dB and BW=10%, noise and LAD thresholds. $N = 64$ samples.

converter (ADC). All the signal processing was performed digitally. Spectrum analyzer was connected to a computer. Instrument Control Toolbox was used to connect Matlab to the spectrum analyzer. This enabled direct results analysis. Six measurements at six different frequency areas were performed from 10 MHz to 39 GHz. Considered frequency ranges were 10-110 MHz, 1-1.1 GHz, 2.5-2.6 GHz, 9-9.1 GHz, 17-17.1 GHz and 39-39.1 GHz. The parameters are presented at Table I. There were 1601 frequency points and 1 000 or 10 000 sweeps in time domain. Energy of the samples was measured, i.e., $|x_i|^2$. The internal noise level of spectrum analyzer was measured in two ways. In the first way, internal noise level was measured when the 50 ohm wideband load was connected to the input of the spectrum analyzer (cases a-d). In the second way, broadband antenna was connected to the input. In this way, noise level is caused by the analyzer internal noise and noise coming from the antenna (cases e and f).

IV. Noise Measurement Results

Results were analyzed using Matlab simulation software. The purpose was to study the statistical properties of noise in different frequency areas, and performance of the LAD method in the presence of measured noise. As a point of comparison, theoretical zero mean Gaussian distributed noise generated from Matlab simulation software was used. Matlab-generated noise was used because Matlab is widely used in the computer simulations, and the performance of the LAD method has already been studied in the presence of Matlab-generated noise. In this way, these measurement results are directly comparable to the earlier results. Energy of those samples was considered, so the used Matlab-generated noise followed chi-squared distribution. Because of the different scales between the measured and simulated energies, energies were normalized.

TABLE I: Measurement parameters. In all cases there were 1601 frequency data points.

Case	Frequency Range	Bandwidth	Sweeps	Antenna
a	10 – 110 MHz	100 MHz	10 000	No
b	1 – 1.1 GHz	100 MHz	10 000	No
c	17 – 17.1 GHz	100 MHz	10 000	No
d	39 – 39.1 GHz	100 MHz	1 000	No
e	2.5 – 2.6 GHz	100 MHz	1 000	Yes
f	9 – 9.1 GHz	100 MHz	1 000	Yes

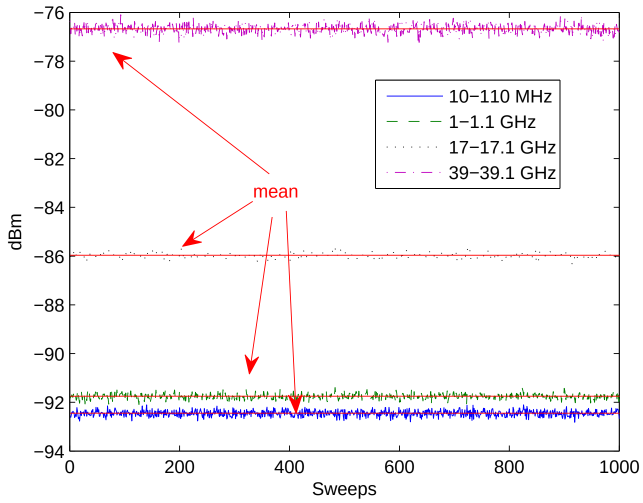


Figure 4: Measured noise energy at different sweeps at different frequency areas.

A. Noise level

From Fig. 4 can be seen that the measured noise levels [dBm] vary a lot at different frequency levels. For example, there is about 15 dB difference between 10-110 MHz and 39-39.1 GHz areas. Thus, adaptive method that is able to estimate the noise level is required when operating at different frequency areas. Instead, methods that use fixed thresholds are not able to operate in all frequency areas without measuring the noise level and defining used threshold based on that information.

B. Histogram

Figs. 5 – 8 present the elements of data into 10 bars that are equally spaced. Number of elements in each container is presented. Cases a, d and e are presented. Therein, the number of elements in each bar is presented. This describes the distribution of energies. Number of time domain sweeps is in y-axis. In Fig. 5, Matlab-generated chi-square distributed noise is used as a reference. For example, first bar consists of about 950 of total 1000 samples. It can be seen that the shapes of histograms are almost equal, so the energies are almost equally-type distributed.

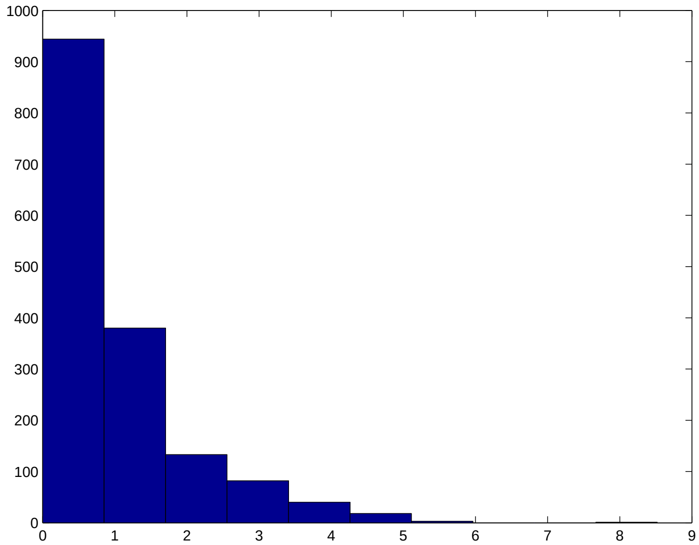


Figure 5: Histogram for Matlab noise.

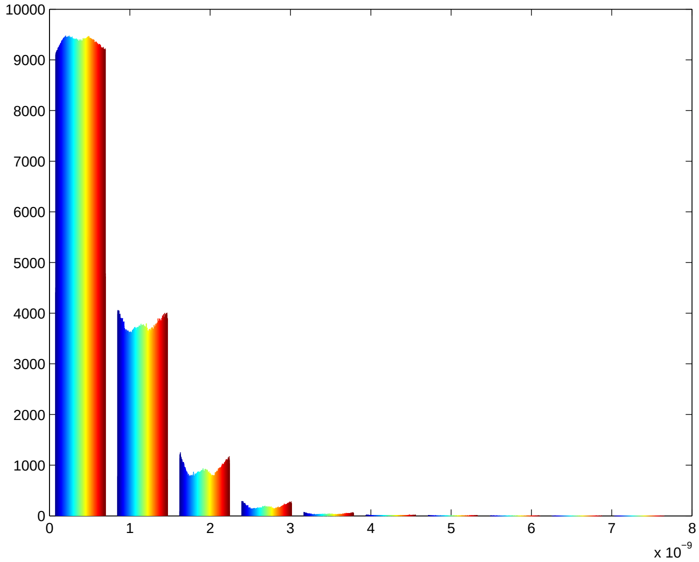


Figure 6: Histogram for 10-110 MHz noise.

C. Probability plot and CDF

The performance of the LAD method depend on the distribution of the noise. In the definition of the LAD method it is assumed that the noise is Gaussian, so variable $|x|^2$ (=energy of samples) follows the chi-squared distribution. Here, it is studied how well the measured noise follows that same distribution, i.e., is there differences between the simulated and measured noise. Fig. 9 presents the probability plots and Fig. 10 presents a plot of the cumulative distribution function (CDF) for the data in the vector x. Empirical CDF (=$F(x)$) can be defined as the proportion of x values less than, or equal, to x. Matlab-generated chi-squared noise was used as a point of comparison. From Figs. 9 and 10 can be seen that there is no differences between the probabilities and CDFs between the measured cases a-f and the Matlab-generated noise.

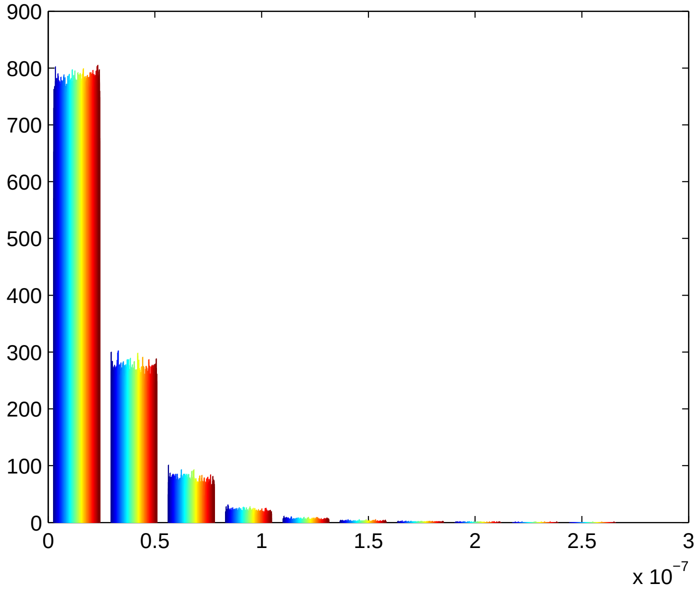


Figure 7: Histogram for 39-39.1 GHz noise.

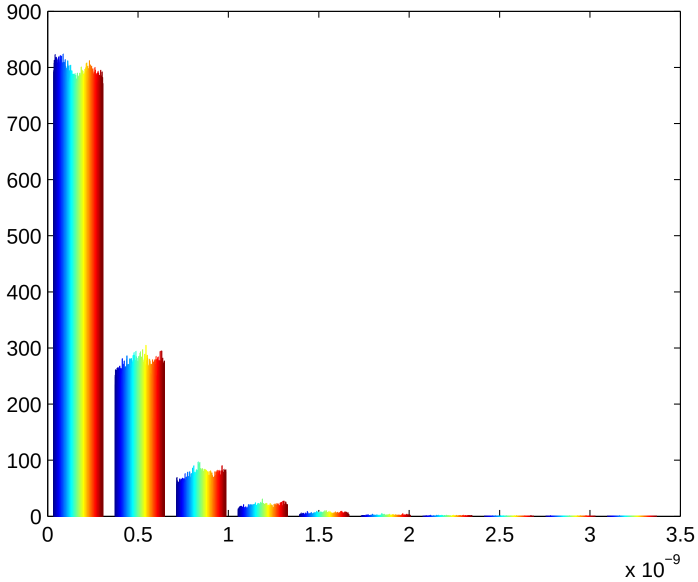


Figure 8: Histogram for 2.5-2.6 GHz noise.

D. Analysis

In this section, the goal is to investigate how the noise at difference frequency areas affect to the probability of detection P_d and probability of false alarm P_{fa} of the LAD method. Also here, Matlab-generated noise is used as a reference. The goal is

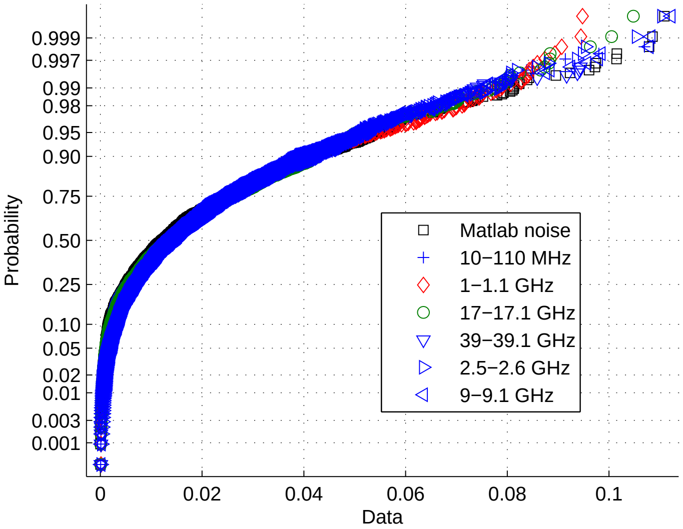


Figure 9: Probability plots for Matlab noise and cases a-f.

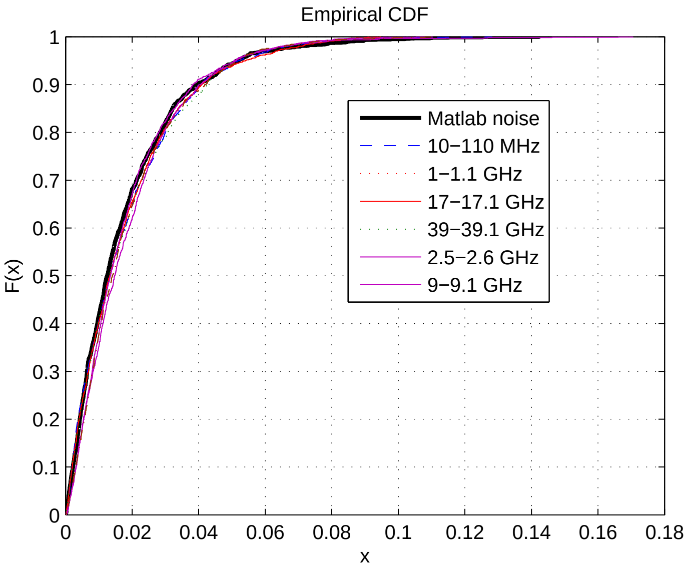


Figure 10: CDF plots for Matlab noise and cases a-f.

not to study the performance of the LAD method, which has already been done, see, for example, [16] and references therein. Here, the purpose is to find out does the measured noise cause any performance degradation compared to Matlab-generated noise. In Fig. 11, P_d vs. SNR is presented. Narrowband (0.3% of the studied bandwidth) theoretical information signal was used as a detected signal. The used LAD threshold parameters were 6.9 (upper) and 2.66 (lower) [16]. It can be noticed that P_d values are approximately on the same level. Note that using smaller upper threshold parameter, signal is found at 0 dB, but there will be more falsely detected signals.

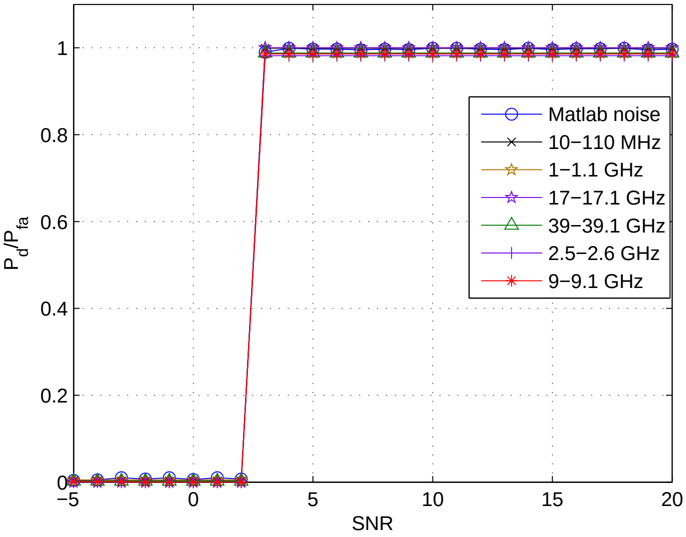


Figure 11: Probability of detecting the signal vs. SNR.

TABLE II: Achieved P_{fa} values. Desired $P_{FA,DES} = 0.01$

Case	Frequency Range	P_{fa}
Matlab noise	-	0.0132
a	$10 - 110$ MHz	0.0064
b	$1 - 1.1$ GHz	0.0062
c	$17 - 17.1$ GHz	0.0061
d	$39 - 39.1$ GHz	0.0072
e	$2.5 - 2.6$ GHz	0.0070
f	$9 - 9.1$ GHz	0.0070

Probability of false alarm results are presented in Table II in the noise-only case. The LAD thresholds were selected so that the desired false alarm probability $P_{FA,DES} = 0.01$, i.e., the upper and lower threshold parameters were 4.6. It means that when there is only noise present, $0.01 = 1\%$ of the samples is above the threshold. Here, 1% corresponds to 16 samples. There is some difference between the desired noise $P_{FA,DES}$ and measured noise P_{fa} values, that is, the measured ones are slightly lower than the desired one. However, the difference is only about 0.003, i.e., 5 samples out of a total of 1601 samples. Performance differences are mainly caused by implementation restrictions of hardware. Noise properties in the analog part of spectrum analyzer may slightly vary in different frequency ranges. In addition, quantization noise affects to the noise properties.

V. Conclusion

Noise measurements were performed at several frequency areas between 10 MHz and 39 GHz. The goal was to study the statistical properties of measured noise in different frequency areas. The measurement results depend on the used equipment. Measured noise characteristics were analyzed and compared also with Matlab-generated noise. It was noticed that as the probability plots were almost equal, there was a great difference between the noise levels. Thus, adaptive spectrum sensing is needed. The LAD spectrum sensing method that is based on the assumption that the noise is Gaussian was studied under the measured noise. It was noticed that the noise had only small effect to the probability of detection and probability of false alarm.

ACKNOWLEDGMENT

The research of Johanna Vartiainen was funded by the Academy of Finland.

REFERENCES

[1] V. Chakravarthy, A. Shaw, M. Temple, and J. Stephens, “Cognitive radio - an adaptive waveform with spectral sharing capability,” in IEEE Wireless Commun. and Networking Conf., New Orleans, LA, USA, Mar.13–17 2005, pp. 724–729.

[2] S. Haykin, “Cognitive radio: Brain-empowered wireless communications,” IEEE Journal in Selected Areas in Comm., vol. 23, no. 2, Feb. 2005, pp. 201–220.

[3] J. Mitola III and G. Q. M. Jr., “Cognitive radio: making software radios more personal,” IEEE Pers. Commun., vol. 6, no. 4, 1999, pp. 13–18.

[4] S. N. Shankar, C. Cordeiro, and K. Challapali, “Spectrum agile radios: Utilization and sensing architectures,” in DySpAN 2005, vol. 1, Baltimore, USA, Nov. 2005, pp. 160–169.

[5] FCC, “FCC frees up vacant TV airwaves for super Wi-Fi technologies,” Sep. 2010, http://www.fcc.gov [retrieved: May, 2014].

[6] M. Lopez-Benitez and F. Casadevall, “Methodological aspects of spectrum occupancy evaluation in the context of cognitive radio,” European. Trans. Telecommun., vol. 21, no. 8, 2010, pp. 680–693.

[7] M. Wellens, “Empirical modelling of spectrum use and evaluation of adaptive spectrum sensing in dynamic spectrum access networks,” Ph.D. dissertation, RWTH Aachen University, Germany, May. 2010.

[8] J. Vartiainen, J. Lehtomäki, and R. Vuohtoniemi, “The LAD methods in WLAN indoor multipath channels,” in CrownCom 2012, Stockholm, Sweden, Jun. 2012, pp. 344–349.

[9] J. Vartiainen, J. J. Lehtomäki, and H. Saarnisaari, “Double-threshold based narrowband signal extraction,” in VTC 2005, Stockholm, Sweden, May/June 2005, pp. 1288–1292.

[10] H. Saarnisaari, P. Henttu, and M. Juntti, “Iterative multidimensional impulse detectors for communications based on the classical diagnostic methods,” IEEE Trans. Commun., vol. 53, no. 3, March 2005, pp. 395–398.

[11] J. G. Poor, Digital Communications, 3rd ed. New York: McGraw-Hill, 1995.

[12] H. Saarnisaari and P. Henttu, “Impulse detection and rejection methods for radio systems,” Boston, MA, USA, Oct. 2003, cD-rom.

[13] J. Vartiainen, J. J. Lehtomäki, H. Saarnisaari, and M. Juntti, “Analysis of the consecutive mean excision algorithms,” J. Elect. Comp. Eng., 2011.

[14] W. Press, W. Vetterling, S. Teukolsky, and B. Flannery, Numerical Recipes in C, 2nd ed. New York: Cambridge University Press, 1992.

[15] “Agilent,” (2014), http://www.agilent.com [retrieved: May, 2014].

[16] J. Vartiainen, “Concentrated signal extraction using consecutive mean excision algorithms,” Ph.D. dissertation, Acta Univ Oul Technica C 368. Faculty of Technology, University of Oulu, Finland, Nov. 2010, http://jultika.oulu.fi/Record/isbn978-951-42-6349-1 [retrieved: May, 2014].

Ordered Sequential-Superposition Cooperative Spectrum Sensing for Cognitive Radio Networks

Hiep Vu-Van
The School of Electrical Engineering
University of Ulsan
Ulsan, Republic of Korea
Email: vvhiep@gmail.com

Insoo Koo
The School of Electrical Engineering
University of Ulsan
Ulsan, Republic of Korea
Email: iskoo@ulsan.ac.kr

***Abstract*—Cognitive radio (CR) is a promising technology for improving usage of frequency band. In CR network, cognitive radio users (CUs) are allowed to use the bands without interference to operation of licensed users. Reliable sensing information about status of primary user (PU), who is assigned a licensed band, is a pre-requirement for CR network. Cooperative spectrum sensing (CSS) is able to offer an improved sensing reliability compared to individual sensing. However, when the number of CUs is large, the latency and network traffic for reporting sensing results to the Fusion Center (FC) become extremely large, which may result in an extended sensing time and collision in the control channel between Cognitive Users (CUs) and the FC. In this paper, we propose an ordered Sequential-Superposition Cooperative Spectrum Sensing (SSCSS) scheme for faster and more reliable spectrum sensing of CR network. Superposition CSS technique extends sensing time to the reporting slots of other CUs until their round of reporting. The proposed scheme estimates the required number of CUs needed to sense for satisfying the reliability requirement of the system. Furthermore, the scheme decides which CUs (and their orders of polling) will be chosen for the sensing process to maximize performance of the proposed scheme. The simulation results of the proposed scheme show the outstanding performance of the proposed scheme compared with the other conventional CSS.**

Keywords–cognitive radio; ordered sequential cooperative spectrum sensing; superposition cooperative spectrum sensing.

I. Introduction

Nowadays, more bandwidth and higher bit-rates have been required to meet usage demands due to an explosion in wireless communication technology. According to the Federal Communications Commission's spectrum policy task force report [1], the actual utilization of the licensed spectrum varies from 15% to 80%. In some cases, the utilization is only a few percent of the total capacity. Cognitive radio (CR) technology [2] has been proposed to solve the problem of ineffective utilization of spectrum bands. Both unlicensed and licensed users, termed the cognitive radio user (CU) and primary user (PU), respectively, operate in CR networks. In CR network, CUs are allowed to access the frequency assigned to PU when it is free. But CU must vacate the occupied frequency when the presence of PU is detected. Therefore, reliable detection of the PU's signal is a requirement of CR networks.

In order to ascertain the presence of a PU, CUs can use one of several common detection methods, such as matched filter, feature, and energy detection [2][3]. Energy detection is the optimal sensing method if the CU has the limited information about PU's signal (e.g., only the local noise power is known) [3]. In energy detection, frequency energy in the sensing channel is collected in a fixed bandwidth W over an observation time window T to compare with the energy threshold and determine whether or not the channel is utilized. However, the received signal power may fluctuate severely due to multipath fading and shadowing effects. Therefore, it is difficult to obtain reliable detection with only one CU. Better sensing performance can be obtained by allowing some CUs to perform cooperative spectrum sensing [4][5][6].

In CSS, because of the limitations of the control channel, CUs will report their sensing information to the FC one by one. Subsequently, in a CR network with a large number of CUs a very large number of reports will be transmitted through a control channel, which can make the sensing process sluggish and result in overhead traffic in the control channel. In order to solve those problems of CSS, SCSS scheme [8][9] has been proposed. In SCSS, the fusion center (FC) acts as the control center for the operation of CR network. The FC sends the "sensing request" message to CUs when it needs their sensing information, and randomly polls CUs one by one until the condition required to make a global decision is satisfied. The ordered SCSS can improve sensing performance by polling sensing results of CUs according to their order of reliability (i.e., signal-to-noise ratio (SNR) of sensing channel of the CU). The ordered SCSS can efficiently reduce the number of sensing report from the CUs. However, the conventional SCSS uses the same sensing time for all CUs. The superposition CSS [8] can solve this problem of conventional SCSS by extending sensing duration of CUs to the reporting time of other CUs.

In this paper, we propose an ordered sequential-superposition CSS for cognitive radio networks. The proposed scheme estimates the required number of CUs needed to poll for satisfying the reliability requirement of the system. Furthermore, through the proposed scheme we can decide which CUs will be chosen for the sensing process and their orders of polling to maximize sensing performance.

This paper is organized as follows. Section 2 describes and analyses the energy detection method. Section 3 gives a detailed explanation of the ordered sequential-superposition cooperative spectrum sensing scheme. Section 4 introduces simulation models and simulation results of the proposed scheme. Finally, Section 5 concludes this paper.

II. System Model

In this paper, we consider a network consisting of N CUs. In addition, there is one PU occupying the observed band with

ISBN: 978-1-61208-360-5

a specific probability. If the CR network needs the sensing information, the FC will send the “request message” to the selected CUs with their order of polling. When the CU receives the “request message” from the FC, it will perform spectrum sensing (SS) and report sensing result to the FC according to its order of polling.

We assume that all CUs utilize energy detector for SS. Then at the i^{th} sensing interval, the received signal energy $E_j(i)$ of the j^{th} CU is given as:

$$E_j\left(i\right) = \begin{cases} \sum_{k=k_i}^{k_i+M_j-1} \left|n_j\left(k\right)\right|^2, & H_0 \\ \sum_{k=k_i}^{k_i+M_j-1} \left|h_j x\left(k\right) + n_j\left(k\right)\right|^2, & H_1 \end{cases} \tag{1}$$

where H_0 and H_1 correspond to the hypotheses of the absence and presence of the PU signal, respectively, $x(k)$ represents the signal transmitted from the PU, h_j denotes the amplitude gain of the channel, and $n(k)$ is the additive white Gaussian noise, $M_j = t_{sj} f_s$ is the number of samples over a sensing interval, t_{sj} is sensing time, f_s is sensing bandwidth and k_i is the time slot at which the i^{th} sensing interval starts.

In conventional CSS, when a CU sends sensing results to the FC, others will keep silent as shown in Fig. 1. In this case, all CUs have the same sensing time such that $t_{s1,C} = t_{s2,C} = ... = t_{sN,C} = t_s$. On the other hand, superposition CSS extends the sensing time of CUs to the reporting time of other CUs as shown in Fig. 2.

When M_j is relatively large (e.g., $M_j > 200$), E_j can be well approximated as a Gaussian random variable under both hypotheses as follows [7]:

$$\begin{array}{l} N\left(\mu_{j,H_0} = M_j, \sigma_{j,H_0}^2 = 2M_j\right) \\ N\left(\mu_{j,H_1} = M_j\left(\gamma_j + 1\right), \sigma_{j,H_1}^2 = 2M_j\left(2\gamma_j + 1\right)\right) \end{array} \tag{2}$$

where $N(.)$ is Gaussian distribution, μ_{j,H_0} and μ_{j,H_1} are the mean of E_j under H_0 and H_1 hypothesis, respectively, σ_{j,H_0}^2 and σ_{j,H_1}^2 are the variance of E_j under H_0 and H_1 hypothesis, respectively, γ_j is SNR in the sensing channel between the j^{th} CU and the PU.

The local decision of the j^{th} CU at the i^{th} sensing interval can be made as the following rule:

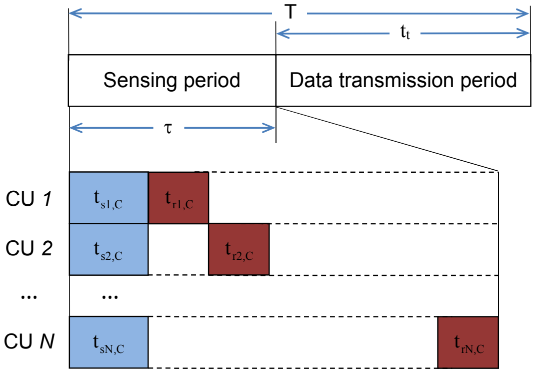


Figure 1. The time frame of conventional cooperative spectrum sensing

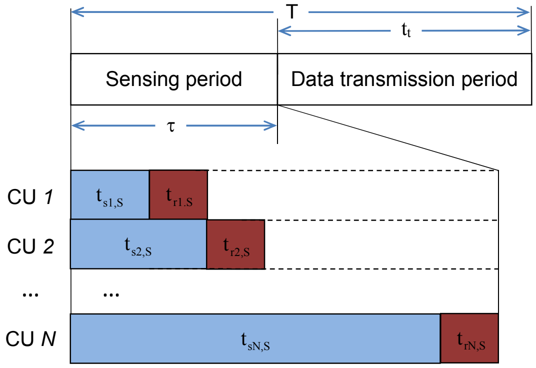


Figure 2. The time frame of superposition cooperative spectrum sensing

$$\begin{cases} G_j\left(i\right) = 1, \text{ if } E_j\left(i\right) \geq \lambda_j \\ G_j\left(i\right) = 0, \text{ otherwise} \end{cases} \tag{3}$$

where λ_j is the threshold for hard local decision of the j^{th} CU.

The average probability of detection and the average probability of false alarm of the j^{th} CU are given, respectively, by [11].

$$\begin{aligned} P_{d,j} &= \Pr\left(G_j\left(i\right) = 1 \middle| H_1\right) \\ &= Q_u\left(\sqrt{2\gamma_j}, \sqrt{\lambda_j}\right), \end{aligned} \tag{4}$$

$$\begin{aligned} P_{f,j} &= \Pr\left(G_j\left(i\right) = 1 \middle| H_0\right) \\ &= \frac{\Gamma\left(M_j, \frac{\lambda_j}{2}\right)}{\Gamma\left(M_j\right)}, \end{aligned} \tag{5}$$

where $\Gamma(a,x)$ is the incomplete gamma function which is given by $\Gamma(a,x) = \int_x^\infty t^{a-1} e^{-t} dt$, $\Gamma(a)$ is the gamma function, $Q_{M_j}(a,b)$ is the generalized Marcum Q-function which is given by $Q_{M_j}(a,x) = \frac{1}{a^{M_j-1}} \int_x^\infty t^{M_j} e^{-\frac{t^2+a^2}{2}} I_{M_j-1}(at) dt$, and $I_{M_j-1}(.)$ is the modified Bessel functions of the first kind and order $(M_j - 1)$.

With the requirement value of probability of detection, $P_{d,j}^*$, probability of false alarm can be calculated as follows:

$$P_{f,j}\left(P_{d,j}^*\right) = Q\left(\sqrt{2\gamma_j + 1} Q^{-1}\left(P_{d,j}^*\right) + \sqrt{M_j}\gamma_j\right) \tag{6}$$

We define the reliability of CU as probability of false alarm $P_{f,j}\left(P_{d,j}^*\right)$. If all CUs have the same $P_{d,j}^*$ and M_j, the CU with lower value of $P_{f,j}\left(P_{d,j}^*\right)$ will be higher reliability.

III. The Ordered Sequential-Superposition Cooperative Spectrum Sensing Scheme

In conventional ordered based SCSS, the highest reliability CU (the CU with the highest SNR of sensing channel) should be polled first for fast SS. However, this technique gives good performance only for CSS with the same sensing time for all CUs. In this paper, we propose an ordered sequential-superposition CSS for cognitive radio network in which the the set of CUs will be selected to perform SS and each of selected CU will be assigned a suitable sensing time for the best sensing performance of sensing process.

In the initial stage, a requirement number of CUs, p, which is needed to perform SS, will be selected as $0 < p < N$. After that, FC will choose the set of p highest reliability CUs, $\Omega = [\mathrm{CU}_1, \mathrm{CU}_2, ..., \mathrm{CU}_\mathrm{p}]$. Set Ω is sorted according to the increasing order of reliability that is CU_1 is the lowest reliable CU and CU_p is the highest reliable CU. The CUs included in set Ω will be required to sense the signal from the PU.

We assume that all CUs have the same reporting time such that $t_{r1,S} = t_{r2,S} = ... = t_{rN,S} = t_r$. Then the sensing time for p CUs will be given as follows:

$$\begin{array}{l}
t_{s1,S} = t_s \\
t_{s2,S} = t_{s1,S} + t_r = t_s + t_r \\
t_{s3,S} = t_{s2,S} + t_r = t_s + 2t_r \\
... \\
t_{sp,S} = t_{sp-1,S} + t_r = t_s + (p-1)\,t_r
\end{array} \tag{7}$$

This means that the CU, who firstly reports sensing information to the FC, will have the shortest sensing time $t_{s1,S} = t_s$ and the CU, who is the last CU reporting sensing information to the FC, will have the longest sensing time $t_{sN,S}$. The time frame of the proposed scheme is shown in the Fig. 3.

In order to maximize sensing performance, in the proposed scheme the CU with higher reliability, CU_p, will be assigned to have the longer sensing time $t_{sp,S}$. Subsequently, the highest reliable CU CU_p is required to sense in $t_{sp,S}$ time and is the last CU reporting sensing information to the FC. On the other hand, the lowest reliable CU, CU_1, is required to sense in $t_{s1,S}$ time and firstly reports sensing information to the FC.

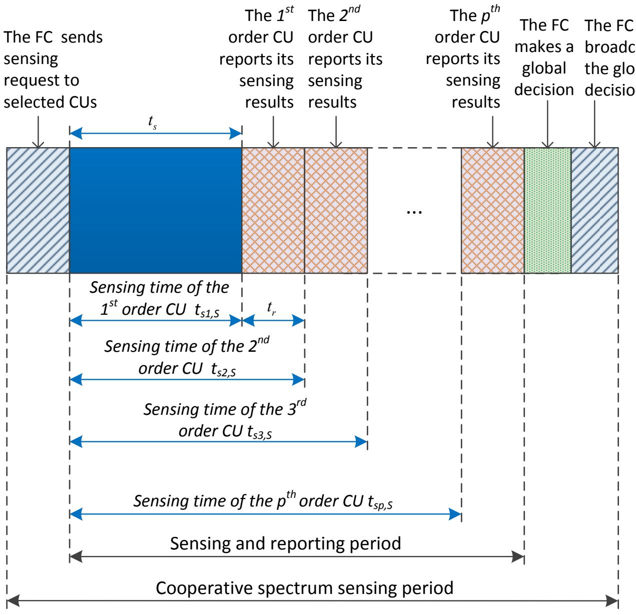

The FC sends sensing request to selected CUs
The 1st order CU reports its sensing results
The 2nd order CU reports its sensing results
The pth order CU reports its sensing results
The FC makes a global decision
The FC broadcasts the global decision
t_s
...
Sensing time of the 1st order CU $t_{s1,S}$
t_r
Sensing time of the 2nd order CU $t_{s2,S}$
Sensing time of the 3rd order CU $t_{s3,S}$
Sensing time of the pth order CU $t_{sp,S}$
Sensing and reporting period
Cooperative spectrum sensing period

Figure 3. The time frame of the proposed scheme

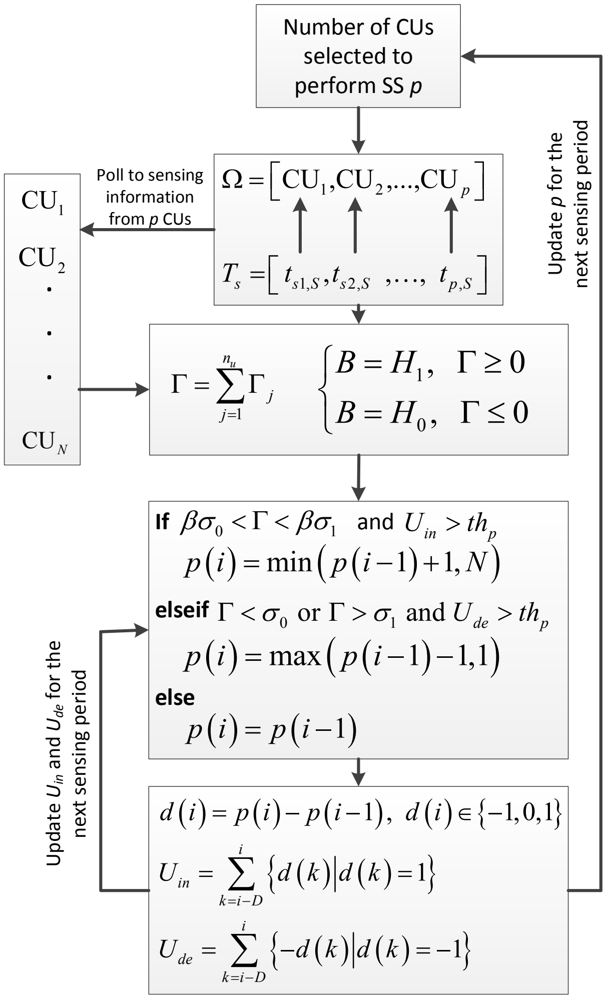

Number of CUs selected to perform SS p
Poll to sensing information from p CUs
$\Omega = [\mathrm{CU}_1, \mathrm{CU}_2, ..., \mathrm{CU}_p]$
$T_s = [t_{s1,S}, t_{s2,S}, ..., t_{p,S}]$
CU_1
CU_2
CU_N
$\Gamma = \sum_{j=1}^{n_u} \Gamma_j$
$\begin{cases} B = H_1, & \Gamma \geq 0 \\ B = H_0, & \Gamma \leq 0 \end{cases}$
If $\beta\sigma_0 < \Gamma < \beta\sigma_1$ and $U_{in} > th_p$
$p(i) = \min(p(i-1)+1, N)$
elseif $\Gamma < \sigma_0$ or $\Gamma > \sigma_1$ and $U_{de} > th_p$
$p(i) = \max(p(i-1)-1, 1)$
else
$p(i) = p(i-1)$
$d(i) = p(i) - p(i-1),\ d(i) \in \{-1, 0, 1\}$
$U_{in} = \sum_{k=i-D}^{i} \{d(k) | d(k) = 1\}$
$U_{de} = \sum_{k=i-D}^{i} \{-d(k) | d(k) = -1\}$
Update U_{in} and U_{de} for the next sensing period
Update p for the next sensing period

Figure 4. Flow-chart of the proposed scheme

In order to start the sensing process, FC will send the “sensing request” message and the order of reporting to the CUs in Ω. When CUs receive the “sensing request” message from the FC, they will sense the signal from the PU until their round of reporting. Each CU will make local decision as Eqn. (3) and report its decision to the FC. At the FC, the accumulated log-likelihood of p CUs will be calculated as [12]:

$$\Gamma = \sum_{j \in \Omega} \Gamma_j \tag{8}$$

where

$$\begin{array}{ll}
\Gamma_j = \log \frac{\mathrm{P_{d,j}}}{\mathrm{P_{f,j}}}, & \text{if } \mathrm{G_j} = 1 \\
\Gamma_j = \log \frac{(1-\mathrm{P_{d,j}})}{(1-\mathrm{P_{f,j}})}, & \text{otherwise.}
\end{array} \tag{9}$$

The global decision about status of the PU signal can be made as:

$$\begin{cases} B = H_1, & \text{if } \Gamma \geq 0 \\ B = H_0, & \text{otherwise} \end{cases} \tag{10}$$

Here, the value of accumulated log-likelihood of p CUs, Γ, is known as reliable level of sensing process. Then we utilize Γ as a criteria to update the required number of CUs for the

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-360-5

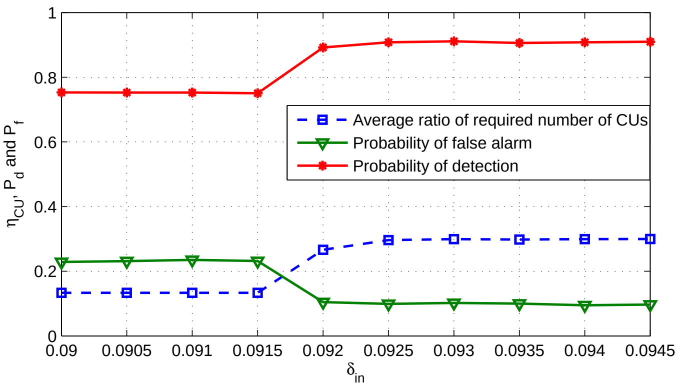


Figure 5. Performance of the proposed scheme versus “*reliable threshold*”

next sensing period. We define σ_1 and σ_0 as the “*reliable thresholds*” of sensing process. Those *reliable thresholds*” can be determined according to requirement of probability of detection, P_d^*, and false alarm, P_f^*, of CR system as [9]:

$$\sigma_0 = \log \frac{\left(1 - P_d^*\right)}{\left(1 - P_f^*\right)} \tag{11}$$

and

$$\sigma_1 = \log \frac{P_d^*}{P_f^*}. \tag{12}$$

We also define the “*fluctuate level*” of p as

$$U_{in} = \sum_{k=i-D}^{i} \left\{ d\left(k\right) | d\left(k\right) = 1 \right\} \tag{13}$$

and

$$U_{de} = \sum_{k=i-D}^{i} \left\{ -d\left(k\right) | d\left(k\right) = -1 \right\}, \tag{14}$$

where U_{in} and U_{de} show the number of times that p is increased and decreased in the considered window size D, and $d(i)$ can be calculated as

$$d\left(i\right) = p\left(i\right) - p\left(i-1\right), \;\; d\left(i\right) \in \left\{-1, 0, 1\right\}. \tag{15}$$

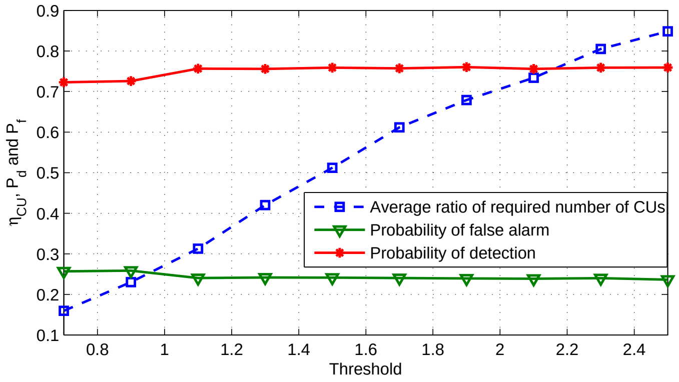


Figure 6. Performance of the conventional ordered SCSS.

TABLE I. Initial Conditions for simulations

Parameter	Initial values
N	30
t_s	1ms
t_r	1ms
th_p	4
n	50000
D	200
U_{in}	5
U_{de}	5
δ_{in}	$\{0.0900, 0.0905, ..., 0.0945\}$
δ_{de}	$\{2.055, 2.070, ..., 2.190\}$

The value of p can be updated at each sensing interval according to values of “*reliable threshold*” and ‘*fluctuate level*” as $p\left(i\right) = \min\left(p\left(i-1\right)+1, N\right)$, if $\beta\sigma_0 < \Gamma < \beta\sigma_1$ and $U_{in} > th_p$, where β is “*adjusting factor*” for “*reliable threshold*” and th_p is threshold for “*fluctuate level*”. If $\Gamma < (2-\beta)\sigma_0$ or $\Gamma > (2-\beta)\sigma_1$ and $U_{de} > th_p$, the value of p will be updated as $p\left(i\right) = \max\left(p\left(i-1\right)-1, 1\right)$. Otherwise, the value of p will be kept the same to that one of the previous sensing interval.

The flow-chart of the proposed scheme is shown in Fig. 4

IV. Simulation Results

In this section, simulation results of the proposed scheme and conventional SCSS with ordered and randomly polling are provided. The network includes 30 CUs with SNR of sensing channel varying from -14dB to -43dB and $t_s = t_r = 1$ms. In order to evaluate the performance in terms of reducing required number of CUs performing SS, we define η_{CU} as average ratio of required number of CUs,

$$\eta_{CU} = \frac{\sum\limits_{i=1}^{n} p\left(i\right)}{nN} \tag{16}$$

where n_i is number of total sensing intervals.

The parameters for simulation are shown in Table I, where the “*reliable thresholds*” are considered as $\delta_{in} = -\beta\sigma_0 = \beta\sigma_1$ and $\delta_{de} = -(2-\beta)\sigma_0 = (2-\beta)\sigma_1$. Fig. 5 shows probability of detection, probability of false alarm and average ratio of required number of CUs, η_{CU}, of the proposed scheme, respectively.

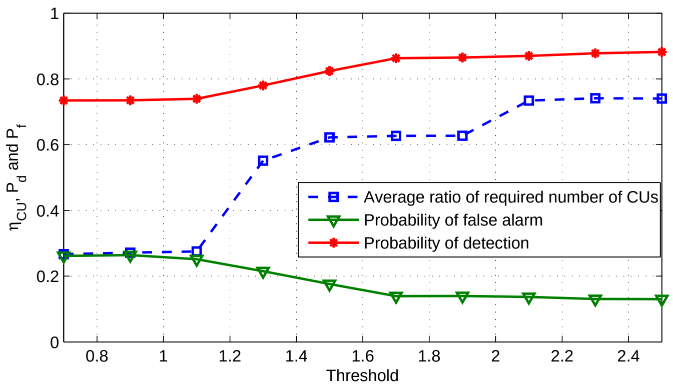


Figure 7. Performance of the conventional randomly polling SCSS.

The performance of reference schemes, conventional SCSS with ordered and randomly polling, are shown in Figs. 6 and 7, respectively. Both schemes consider superposition for assigning sensing time for each CU. The ordered SCSS polls sensing information from CUs according to their values of SNR in the sensing channel; the CUs with higher SNR will be polled sooner than the CUs with lower SNR. The randomly polling SCSS randomly choose the CUs to poll sensing information.

From Figs. 5, 6 and 7, it can be observed that the proposed scheme has the best performance. When $\eta_{CU} = 0.3$, the proposed scheme obtains the sensing performance of $P_d = 0.9$ and $P_f = 0.1$; however, the randomly polling SCSS and ordered SCSS obtains sensing performance of $P_d = 0.75$ and $P_f = 0.25$. The randomly polling SCSS can get the similar sensing performance to that of the proposed scheme when its required number of CUs is two time higher (i.e., $\eta_{CU} = 0.6$) than that of the proposed scheme. For the conventional ordered SCSS, most of high reliable CUs are polled to achieve good sensing performance at $P_d = 0.75$ and $P_f = 0.25$, and the performance cannot be improved even when the number of polled CUs is increased.

V. Conclusion

In this paper, an ordered sequential-superposition CSS is proposed for fast SS. The proposed scheme shows the algorithm to determine how many and which CUs are needed to sense the signal from PU and their corresponding sensing time for superposition CSS. The simulation results prove that the proposed scheme significantly improves performance of sensing process and can reduce 50% of required number of CUs to achieve the similar sensing performance to conventional SCSS.

Acknowledgement

This work was supported by the KRF funded by the MEST (NRF-2012R1A1A2038831 and NRF-2013R1A2A2A05004535).

References

[1] Federal Communications Commission, Spectrum Policy Task Force, Rep. ET Docket no. 02-135, 2002.

[2] Y. Hur et al., “A wideband analog multi-resolution spectrum sensing (MRSS) technique for cognitive radio (CR) systems”, in Proc. IEEE Int. Symp. Circuit and System, Greece, pp.4090-4093, 2006.

[3] A. Sahai, N. Hoven and R. Tandra, “Some fundamental limits on cognitive radio”, in Proc. Allerton Conf. on Communications, control, and computing, Monticello, 2004, pp.1-11.

[4] G. Ganesan and Y. G. Li, “Cooperative spectrum sensing in cognitive radio networks”, in Proc. IEEE Symp. New Frontiers in Dynamic Spectrum Access Networks (DySPAN05), Baltimore, USA, 2005, pp.137-143.

[5] S. M. Mishra, A. Sahai and R. W. Brodersen, “Cooperative sensing among CRs,” IEEE International Conf. Commun., vol. 4, pp.1658-1663, 2006.

[6] R. Deng, J. Chen, C. Yuen, P. Cheng and Y. Sun, “Energy-Efficient Cooperative Spectrum Sensing by Optimal Scheduling in Sensor-Aided Cognitive Radio Networks,” IEEE Transactions on Vehicular Technology, vol.61, no.2, 2012, pp.716-725.

[7] J. Ma and Y. Li, “Soft Combination and Detection for Cooperative Spectrum Sensing in Cognitive Radio Networks,” Global Telecommunications Conference, GLOBECOM, 2007, pp.3139-3143.

[8] R. Chen, J. Park and K. Bian, “Robust Distributed Spectrum Sensing in Cognitive Radio Networks,” The 27th Conference on Computer Communications (INFOCOM), IEEE., 2008, pp.1876-1884.

[9] Q. Zou, S. Zheng and A. H Sayed, “Cooperative Sensing via Sequential Detection,” IEEE Transactions on Signal Processing, vol.58, no.12, 2010, pp.6266-6283.

[10] J. Jin, H. Xu, H. Li and C. Ren, “Superposition-based cooperative spectrum sensing in cognitive radio networks,” Computer Application and System Modeling (ICCASM), 2010 International Conference on , vol.4, 2010, pp.V4-342,V4-346.

[11] F. F. Digham, M.-S. Alouini and M. K. Simon, “On the energy detection of unknown signals over fading channels,” in Proc. IEEE Int. Conf. Commun., Anchorage, AK, USA, 2003, pp.3575-3579.

[12] Z. Chair and P. K. Varshney, “Optimal data fusion in multiple sensor detection systems,” IEEE Trans. Aero. Electron. Syst., 1986, pp.98-101.

Automatic Floor Map Construction for Indoor Localization

Xin Luo, Albert Kai-sun Wong, Mu Zhou, Xuning Zhang, and Chin-Tau Lea
Electronic and Computer Engineering Department
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
Emails: xluo@connect.ust.hk, eealbert@ust.hk, zhoumu@cqupt.edu.cn, eexuning@ust.hk, and eelea@ece.ust.hk.

***Abstract*—Existing indoor localization systems based on Wi-Fi Received Signal Strength (RSS) fingerprinting often assume the knowledge of a map of the coverage area and involve a tedious manual survey process at a set of sample locations along this map. In this paper, we describe an automatic graphical floor map and radio fingerprints generation system, called the Intelligent Mobility Mapping System (IMMS), which applies the concepts of crowd-sourcing and Simultaneous Localization and Mapping (SLAM) to construct a floor map in support of indoor people localization and tracking. IMMS makes use of high similar patterns in crowd-sourced traces of RSS measurements to identify location segments in the coverage area and to construct a graphical floor map. With IMMS, the elaborate off-line manual data collection process is eliminated.**

Keywords*–*Wi-Fi SLAM; Indoor Localizations; Graph Theory; Mobility Mapping.

I. INTRODUCTION

Indoor localization based on radio signals has been a focus of research for over 10 years. Kong et al. [1] study the use of CDMA2000 pilot signals to record the fingerprints for radio map construction. The fingerprinting approach for indoor localization typically requires a time consuming off-line survey process for building a radio map. Various proposals have been made to reduce the complexity of this process. Ouyang et al. [2], proposed to use unlabeled Wi-Fi Received Signal Strength (RSS) data to enhance the radio map created by labeled survey data. Zhang et al. [3], used a pedestrian mobility model based on knowledge of the physical indoor map to enhance localization.

Recently, the new concept of Simultaneous Localization and Mapping (SLAM) has been developed with the objective of minimizing the off-line survey effort [4]. A SLAM system gathers location and mapping information at the same time, and requires only a very short time for off-line survey. Typically, SLAM records a user's 'footprint' using Microelectromechanical System (MEMS) devices such as accelerometer, gyroscopes and magnetometers on a mobile device, and labels the RSS measurements with this footprint information to construct the mobility maps. Then, the system can locate users on this constructed indoor map in the on-line stage. For example, the system proposed by Shin et al. [5] constructs a floor plan of a building by integrating the number of walking steps (from pedometer), the walking orientation (from magnetometer), and the RSS values recorded with user movements. Zhou et al. [6] use only Wi-Fi RSS measurements from one or multiple individuals walking around the coverage. Similar RSS measurements are clustered and aligned to construct a map for the coverage area.

A. Main contributions

In this paper, we present a system called the Intelligent Mobility Mapping System (IMMS), which is a crowd-sourced system that sporadically collects traces of RSS measurements from users moving around the indoor coverage area as they carry on their daily routines. IMMS automatically creates graphical floor maps to support Wi-Fi RSS-based indoor localization and tracking. There are three modules in IMMS. The first module is designed to facilitate the subsequent processes by various data pre-processing methods. The second module is designed to find the highly similar pieces of measurements in traces. IMMS estimates similarities of measurements in different traces by correlations. Then, the resemblant measurements in different RSS traces are clustered as High Cross-Trace Correlation Patterns (HCP), and the intersections of these HCPs are used to segment traces into Atomic Location Segments (ALSs). The third module is designed to construct the draft graphical floor map and radio map. The floor map is a simple planar embedding of a drawing that represents the interconnections of ALSs.

Section II overviews the system framework. Section III describes the location segment recognition algorithm and Section IV explains the graph drawing procedures.

B. Notations

Notations to be used in this paper are first summarized in Table I.

TABLE I. IMPORTANT NOTATIONS

Symbol	Meaning
$\mathbf{R^l}$	l^{th} trace
$\nu_\mathbf{i}^\mathbf{l}$	i^{th} measurement in l^{th} trace
$\mathbf{S}$	Raw data matrix
$\mathcal{N}$	Number of Traces in $\mathbf{S}$
Υ	Number of measurements in $\mathbf{S}$
$\mathcal{M}$	Number of hearable APs in the target area
$\mathbf{C}$	Cross-Trace Correlation (CTC) matrix
$\mathbf{C^{\{f,g\}}}$	Submatrix of CTC matrix corresponds to $\mathbf{R^f}$ and $\mathbf{R^g}$
$C_{i,j}^{\{f,g\}}$	A element in $\mathbf{C^{\{f,g\}}}$
$\mathbf{Q}$	Quantized matrix of $\mathbf{C}$
x_i	Row breaking points
y_i	Column breaking points
$\mathbf{G}$	Geometric map
$\mathbf{U}$	Set of unique vertices/ALS of graph G
$\mathbf{E}$	Set of unique edges of graph G
$d(*)$	Direction of an edge
I	Indication matrix of endpoints

II. SYSTEM OVERVIEW

During the data collection phase, traces of RSS measurements are recorded from mobile phones when users move around the coverage area as they conduct their daily activities

ISBN: 978-1-61208-360-5

indoors. For a user, the RSS measurements are collected at a regular time interval (1 read/sec by default) when movement is detected. Data collection stops when the user ceases moving. Each measurement is a sample vector that contains the measured RSSs from a set of hearable Wi-Fi APs. The recorded traces are uploaded to IMMS in cloud.

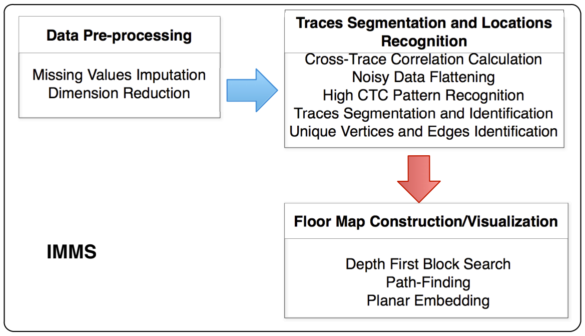

Data Pre-processing
Missing Values Imputation
Dimension Reduction
Traces Segmentation and Locations Recognition
Cross-Trace Correlation Calculation
Noisy Data Flattening
High CTC Pattern Recognition
Traces Segmentation and Identification
Unique Vertices and Edges Identification
Floor Map Construction/Visualization
Depth First Block Search
Path-Finding
Planar Embedding
IMMS

Figure 1. The System Framework

IMMS compares the unlabeled measurements in each pair of traces to see if there are highly similar ones. It is expected that measurements in two traces would exhibit similar distributions if they were collected at nearby places. A string of similar measurements between two traces form what we call a High Correlation Pattern (HCP). Intersecting the HCPs form by a given trace with all other traces allows us to segment HCPs into what we call ALS, which we expect would represent corridor sections between intersections in the physical environment. By observing how the ALSs are connected in the traces, we can then construct a 2D graphical floor map. The architecture of IMMS is shown in Fig.1. It contains three modules: data pre-processing, traces segmentation and locations recognition, and floor map.

The data pre-processing module includes missing value imputation and dimension reduction. Because of the limited coverage of each AP and the random variation of the RSS signals, most of the recorded RSS values are zero. In order to insure the accuracy of the similarity evaluation, we need to find a way to impute these zero, or missing values. Based on recommendation provided by Ouyang et al. [7], we impute missing values with a number that is smaller than the minimum collected RSS measurements. Moreover, because the total number of APs in the coverage area can be very large [6], it is desirable to reduce the data dimensionality to reduce computation complexity. We apply Principal Component Analysis (PCA) [8] to reduce the measurement dimensionality.

The details of the second module ,traces segmentation and locations recognition module, and the third module, floor map construction module, will be given in Section III and Section IV, respectively.

III. Traces Segmentation and Locations Recognition

Our approach is based on the premise that the physical space can be modeled as an interconnection of corridor segments that we call ALSs. Users tend to traverse an ALS in its entirety, and in one of two directions. Each recorded trace of Wi-Fi measurements reflects the movement of a user through a number of ALSs.

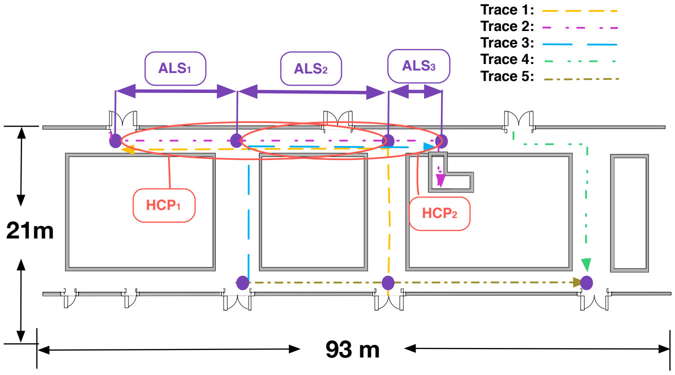


Figure 2. Example of Traces, HCPs and ALSs

The objective of traces segmentation is to identify all ALSs in the physical environment using traces from crowdsourcing. The traces segmentation process starts with the recognition of any high similarity pattern, which we call High Cross Trace Correlation (CTC) Pattern, that may exist in any given pair of traces. A High CTC Pattern (HCP) reflects a sequence of overlapping ALSs between two traces. From the starting points and ending points of all HCPs found in many trace pairs, we can identify all the Breaking Points (BPs) which separate the ALSs in all the traces. In other words, we identify all the individual ALSs by intersecting all the HCPs. An example is shown in Fig. 2. Computation of similarity is based on the dimension-reduced measurements after data pre-processing.

A. *Cross-Trace Correlation (CTC)*

We measure the Cross-Trace Correlation (CTC) by the Pearson product-moment [9] correlation coefficient between measurements in two different traces. Assume $\mathcal{N}$ traces are collected. Let $\Re^{\mathbf{f}} = (\nu_1^f, ..., \nu_{n^f}^f)$ be the f^{th} $(f = 1, ..., \mathcal{N})$ trace, where ν_i^f is the $i^{th}(i = 1, ..., n^f)$ measurement in trace $\Re^{\mathbf{f}}$, and n^f is the number of measurements in the trace. Each measurement is a row vector $\nu_i^f = [\nu_{i,1}^f ... \nu_{i,\mathcal{M}}^f]$, where $\nu_{i,j}^f$ is the signal strength from the $j^{th}(j = 1, ..., \mathcal{M})$ AP. The raw data of all $\mathcal{N}$ traces can be kept in a $\Upsilon \times \mathcal{M}$ matrix $\mathcal{S} = [\Re^{\mathbf{1}}\Re^{\mathbf{2}} ... \Re^{\mathcal{N}}]^T$, where $\Upsilon = \sum_{f=1}^{\mathcal{N}} n^f$ is the total number of measurements in the data. After dimension reduction, each measurement is reduced to k dimension, denoted as $\mu_{\mathbf{i}}^{\mathbf{l}}$ ($k-$dimension row vector). The dimension reduced data matrix is $\mathbf{S} = [\mathbf{R}^{\mathbf{1}}\mathbf{R}^{\mathbf{2}} ... \mathbf{R}^{\mathcal{N}}]^T$ ($\Upsilon \times k$ matrix). We define CTC matrix, $\mathbf{C}$, as the matrix containing the sub-matrices $\mathbf{C}^{\{\mathbf{f},\mathbf{g}\}}$, where $f, g = \{1, ..., \mathcal{N}\}$. $\mathbf{C}^{\{\mathbf{f},\mathbf{g}\}}$ contains CTC values of measurements in trace $\mathbf{R}^{\mathbf{f}}$ and $\mathbf{R}^{\mathbf{g}}$, denoted as $C_{i,j}^{\{f,g\}}$.

Definition 1: The CTC value of measurements $\mu_i^f \in \mathbf{R}^{\mathbf{f}}$ and $\mu_j^g \in \mathbf{R}^{\mathbf{g}}$ is

$$C_{i,j}^{\{f,g\}} = corr(\mu_{\mathbf{i}}^{\mathbf{f}}, \mu_{\mathbf{j}}^{\mathbf{g}}) = \frac{1}{\sigma_i^f \sigma_j^g} E[(\mu_{\mathbf{i}}^{\mathbf{f}} - \mathbf{m})(\mu_{\mathbf{j}}^{\mathbf{g}} - \mathbf{m})]. \quad (1)$$

The column mean of the dimension reduced sample matrix $\mathbf{S}$ is $\mathbf{m} = [m_1 ... m_k]$. Then, $m_j = \frac{\sum_{f=1}^{N} \sum_{i=1}^{n^f} \mu_{\mathbf{i},\mathbf{j}}^{\mathbf{f}}}{\Upsilon}$. σ_i^f, σ_j^g are

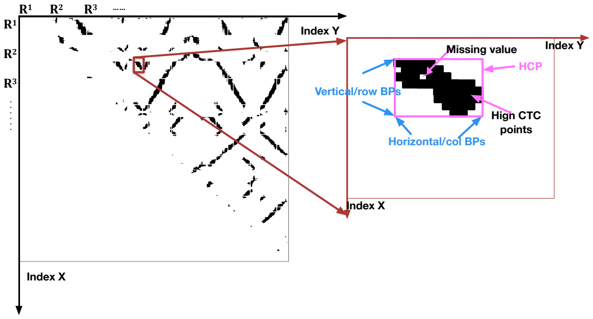


Figure 3. Examples of data structure and HCP

the standard derivation of measurements $\mu_{\mathbf{i}}^{\mathbf{f}}$ and $\mu_{\mathbf{j}}^{\mathbf{g}}$, calculated as $\sigma_i^f = \sqrt{\sum_{j\epsilon k}(\mu_{i,j}^f - m_j)^2}$.

B. High CTC Pattern Recognition (HCPR)

The High CTC Pattern Recognition (HCPR) algorithm is used to identify groups of correlated measurements in each pair of traces. If two measurements, $\mu_{\mathbf{i}}^{\mathbf{f}} \in \mathbf{R}^{\mathbf{f}}$ and $\mu_{\mathbf{j}}^{\mathbf{g}} \in \mathbf{R}^{\mathbf{g}}$ have high CTC value, we assume they are collected at nearby physical locations in the two traces and we expect that:

- $\mu_{\{\mathbf{i+1}\}}^{\mathbf{f}}$ and $\mu_{\{\mathbf{j+1}\}}^{\mathbf{g}}$ will continue to be a high CTC point if the two users are traversing an ALS in the same direction;
- $\mu_{\{\mathbf{i+1}\}}^{\mathbf{f}}$ and $\mu_{\{\mathbf{j-1}\}}^{\mathbf{g}}$ will continue to be a high CTC point if the two users are traversing an ALS in opposite directions.

Thus, the high CTC points should continue in either the southeasterly direction or southwesterly direction until trace $\mathbf{R}^{\mathbf{f}}$ and trace $\mathbf{R}^{\mathbf{g}}$ diverge to two distinct ALSs. We call the point that breaks the continuity of the high CTC points a *breaking point* (BPs), shown in Fig. 3.

The HCPR algorithm first quantizes all the CTC points in the evaluated submatrix to two quantization levels, setting all CTC values bigger than a given threshold $thres$ as 1 and all CTC values lower than $thres$ as 0. The resulting matrix is denoted as $\mathbf{Q}^{\{\mathbf{f},\mathbf{g}\}}$. Then, HCRP groups a contiguous set of high CTC points in the quantized matrix into a high CTC Pattern (HCP).

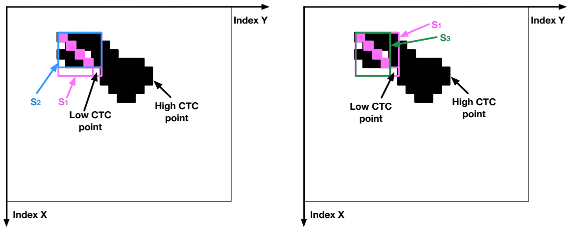
Missing values and random variations in the raw measurement may lead to unwanted breaks in the high CTC pattern. Additionally, the walking speeds of different users are different, and so the height and width of a HCP may be different. We apply the following criteria to determine the boundary of an HCP:

1) HCPR amounts to finding the row indexes and column indexes of each submatrix in Q that encloses a continuous cluster of 1s.

2) There should be a sufficiently large gap between two different HCPs.

3) Small HCPs should be discarded as noise.

Our algorithm for finding HCPs works as follow. We start from the top row of each sub-matrix $\mathbf{Q}^{\{\mathbf{f},\mathbf{g}\}}$ and scan from left to the right for 1s. If none is found, we move to the next row below and scan from left to right again. Once a 1, or a high CTC point, is found, we mark the column and row index of this point as (x, y), and keep searching in the southeasterly or southwesterly direction for 1s. After an HCP is identified, we resume the search for new HCP on the row we left off from. Points covered in a search will be precluded from future search. Starting from a 1, if the point to the southeast is also a 1, then HCPR continues to look for a 1 in the southeast. If a 0 appears, then HCPR calculates the sums of the three sub-matrices indicated in Fig. 4. It compares s_1 and s_2, and s_1 and s_3. If $s_2 = s_1$, it means that there are no new high CTC points is added when the row number is increased (Fig. 4(a)), then the row gap counter: $row_{gap} = row_{gap} + 1$. If $s_2 > s_1$, then one or more new high CTC points are added when the row number is increased. In a similar way, if $s_3 = s_1$, the column gap counter: $col_{gap} = col_{gap} + 1$. If $s_3 > s_1$, then one or more new high CTC points is added because of when the column number is increased (Fig. 4(b)). Once row_{gap} or col_{gap} reaches the defined gap threshold gap_{thres}, HCPR stores the row/column end point of the current HCP, and stops the search of 1s in the southeasterly direction. Next, HCPR starts searching for 1s in the southwesterly direction, using the same search and stop logic as above. Finally, HCPR labels the current HCP using the smallest stored row index and smallest stored column index as well as the largest stored row index and the largest stored column index. The smallest and largest row indexes represent the BPs in trace $\mathbf{R}^{\mathbf{f}}$ and the smallest and largest column indexes represent the BPs in trace $\mathbf{R}^{\mathbf{g}}$. For any possible new HCP between trace $\mathbf{R}^{\mathbf{f}}$ and $\mathbf{R}^{\mathbf{g}}$, HCPR starts searching on row x five entries to the right of the largest column index of the current HCP. The details of HCPR are specified in **Algorithm 1**.



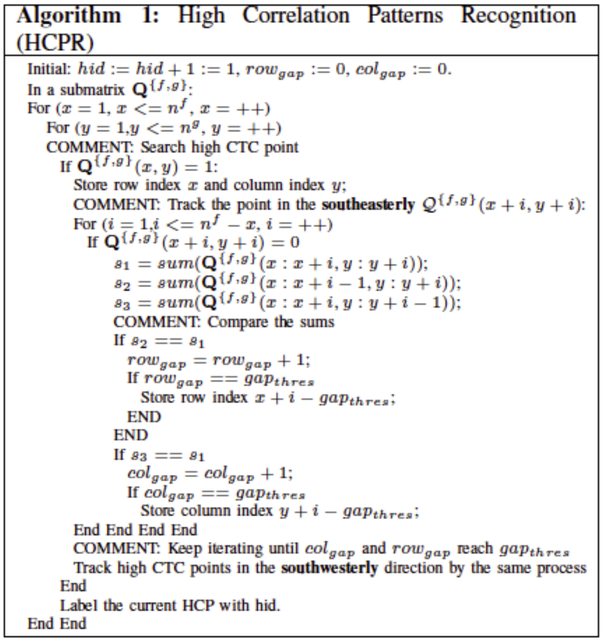
(a) (b)

Figure 4. Example of sub-matrices s_1, s_2, s_3

C. Traces Segmentation and Identification Algorithm

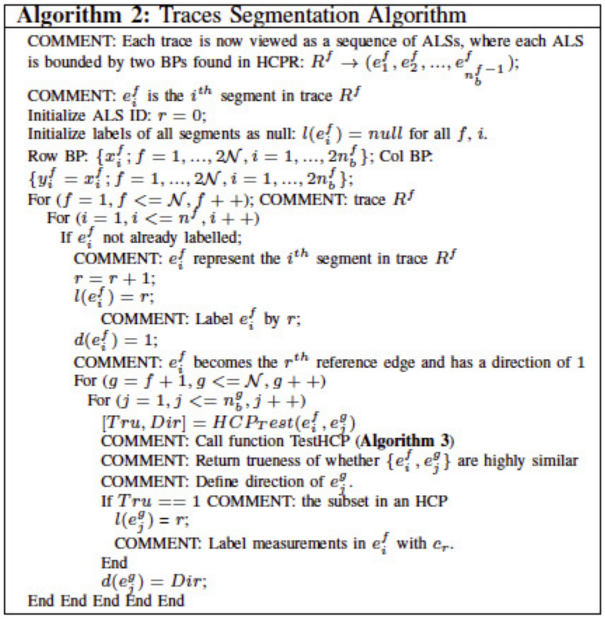
From HCPR, we determine the end points of all HCPs. These end points are the BPs in physical paths at which user paths may diverge. Let $\{x_1^1, ..., x_{n_1}^1, x_1^f, ..., x_{n_f}^f x_1^N, ..., x_{n_N}^N\}$ be the set of row indexes in increasing order which are marked as BPs in the vertical direction of $\mathbf{Q}$, where x_i^f is the i-th BP marked for trace R^f and n_f the number of BPs marked for trace R^f. Likewise, let $\{y_1^1, ..., y_{n_1}^1, y_1^f, ..., y_{n_f}^f y_1^N, ..., y_{n_N}^N\}$ be the set of column indexes in increasing order which are marked as BPs in the horizontal direction of $\mathbf{Q}$. Although $\mathbf{Q}$ is symmetric, because of the way we identify the BPs in HCPR, some x_i^f and y_i^f can become different (empirically most of them are the same).

Our algorithm partitions all the traces using the BPs identified and assigns a unique label (ALS ID) to segments



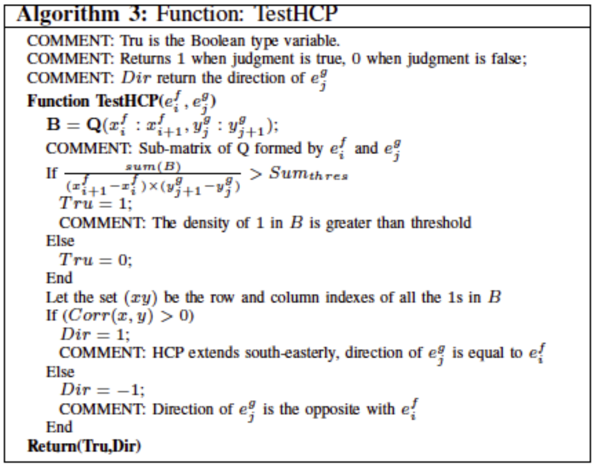
```
Algorithm 1: High Correlation Patterns Recognition (HCPR)
Initial: $hid := hid + 1 := 1$, $row_{gap} := 0$, $col_{gap} := 0$.
In a submatrix $\mathbf{Q}^{\{f,g\}}$:
For ($x = 1$, $x <= n^f$, $x = ++$)
  For ($y = 1$,$y <= n^g$, $y = ++$)
  COMMENT: Search high CTC point
    If $\mathbf{Q}^{\{f,g\}}(x, y) = 1$:
      Store row index $x$ and column index $y$;
      COMMENT: Track the point in the **southeasterly** $Q^{\{f,g\}}(x + i, y + i)$:
      For ($i = 1$,$i <= n^f - x$, $i = ++$)
        If $\mathbf{Q}^{\{f,g\}}(x + i, y + i) = 0$
            $s_1 = sum(\mathbf{Q}^{\{f,g\}}(x : x + i, y : y + i))$;
            $s_2 = sum(\mathbf{Q}^{\{f,g\}}(x : x + i - 1, y : y + i))$;
            $s_3 = sum(\mathbf{Q}^{\{f,g\}}(x : x + i, y : y + i - 1))$;
            COMMENT: Compare the sums
            If $s_2 == s_1$
              $row_{gap} = row_{gap} + 1$;
              If $row_{gap} == gap_{thres}$
                Store row index $x + i - gap_{thres}$;
              END
            END
            If $s_3 == s_1$
              $col_{gap} = col_{gap} + 1$;
              If $col_{gap} == gap_{thres}$
                Store column index $y + i - gap_{thres}$;
      End End End End
      COMMENT: Keep iterating until $col_{gap}$ and $row_{gap}$ reach $gap_{thres}$
      Track high CTC points in the **southwesterly** direction by the same process
    End
    Label the current HCP with hid.
End End
```

Figure 5. High Correlation Patterns Recognition Algorithm



```
Algorithm 2: Traces Segmentation Algorithm
COMMENT: Each trace is now viewed as a sequence of ALSs, where each ALS
is bounded by two BPs found in HCPR: $R^f \rightarrow (e_1^f, e_2^f, ..., e^f_{n_b^f - 1})$;
COMMENT: $e_i^f$ is the $i^{th}$ segment in trace $R^f$
Initialize ALS ID: $r = 0$;
Initialize labels of all segments as null: $l(e_i^f) = null$ for all $f$, $i$.
Row BP: $\{x_i^f; f = 1, ..., 2\mathcal{N}, i = 1, ..., 2n_b^f\}$; Col BP:
$\{y_i^f = x_i^f; f = 1, ..., 2\mathcal{N}, i = 1, ..., 2n_b^f\}$;
For ($f = 1, f <= \mathcal{N}, f + +$); COMMENT: trace $R^f$
  For ($i = 1, i <= n^f, i + +$)
    If $e_i^f$ not already labelled;
      COMMENT: $e_i^f$ represent the $i^{th}$ segment in trace $R^f$
      $r = r + 1$;
      $l(e_i^f) = r$;
          COMMENT: Label $e_i^f$ by $r$;
      $d(e_i^f) = 1$;
      COMMENT: $e_i^f$ becomes the $r^{th}$ reference edge and has a direction of 1
      For ($g = f + 1, g <= \mathcal{N}, g + +$)
        For ($j = 1, j <= n_b^g, j + +$)
          $[Tru, Dir] = HCP_Test(e_i^f, e_j^g)$
          COMMENT: Call function TestHCP (**Algorithm 3**)
          COMMENT: Return trueness of whether $\{e_i^f, e_j^g\}$ are highly similar
          COMMENT: Define direction of $e_j^g$.
          If $Tru == 1$ COMMENT: the subset in an HCP
            $l(e_j^g) = r$;
            COMMENT: Label measurements in $e_i^f$ with $c_r$.
          End
          $d(e_j^g) = Dir$;
End End End End End
```

Figure 7. Trace Segmentation Algorithm



```
Algorithm 3: Function: TestHCP
COMMENT: Tru is the Boolean type variable.
COMMENT: Returns 1 when judgment is true, 0 when judgment is false;
COMMENT: $Dir$ return the direction of $e_j^g$
**Function TestHCP**($e_i^f, e_j^g$)
  $\mathbf{B} = \mathbf{Q}(x_i^f : x_{i+1}^f, y_j^g : y_{j+1}^g)$;
  COMMENT: Sub-matrix of Q formed by $e_i^f$ and $e_j^g$
  If $\frac{sum(B)}{(x_{i+1}^f - x_i^f) \times (y_{j+1}^g - y_j^g)} > Sum_{thres}$
    $Tru = 1$;
    COMMENT: The density of 1 in $B$ is greater than threshold
  Else
    $Tru = 0$;
  End
  Let the set $(xy)$ be the row and column indexes of all the 1s in $B$
  If ($Corr(x, y) > 0$)
    $Dir = 1$;
    COMMENT: HCP extends south-easterly, direction of $e_j^g$ is equal to $e_i^f$
  Else
    $Dir = -1$;
    COMMENT: Direction of $e_j^g$ is the opposite with $e_i^f$
  End
**Return(Tru,Dir)**
```

Figure 8. TestHCP function

in different traces that recognized as the same ALS. The algorithm is described by the pseudo-code in **Algorithm 2** and **Algorithm 3**. In **Algorithm 2**, each trace R^f is considered as a sequence of ALSs. Each segment, e_i^f, is a cluster of RSS measurements between two row BPs x_i^f and $x_i^f + 1$. If a segment has not yet been labelled, we label it with a new ALS ID, and then consider different traces R^g (for $g > f$) and check whether if any ALSs e_j^g in R^g forms an HCP with e_i^f. This checking is via the test function **TestHCP** as described in **Algorithm 2**. Basically, the algorithm checks whether the average number of "1"s in the submatrix of $\mathbf{Q}$ formed by e_i^f and e_j^g is greater than a given threshold. **TestHCP** also determines whether e_j^g is in the same or opposite direction as e_i^f, based on whether the column indexes and row indexes of the "1"s within the sub-matrix is positively or negatively correlated.

As result, we derive from each trace a sequence of ALSs. Each e_i^f is identified with an ALS ID $l(e_i^f) = r$ and marked with directionality $d(e_i^f) = \pm 1$. A sample of labeled ALSs is shown in Fig. 6.

D. Unique Vertices and Edges Identification Algorithm

Next, we proceed to identify and label all the unique vertices that connect the ALS's. Let s_r represent the starting vertex and t_r the terminating vertex of ALS r in the reference direction. Then, we can represent each segment e_i^f in a trace by a tuple of two vertices as follow:

Assume $l(e_i^f) = r$. If $d(e_i^f) = +1$, then $e_i^f = (s_r, t_r)$; if $d(e_i^f) = -1$, then $e_i^f = (t_r, s_r)$.

Then, we can identify vertices that are the same by examining the sequence of segments in all traces, using an indicator $\mathbf{I}$ to record the result as follows: For trace $\mathbf{R^f}$, if $l(e_i^f) = r$, $l(e_{i+1}^f) = r'$, we set:

$$\begin{cases} \mathbf{I}(t_r, s_{r'}) = 1 & \text{if} \quad d(e_i^f) = +1, d(e_{i+1}^f) = +1; \\ \mathbf{I}(t_r, t_{r'}) = 1 & \text{if} \quad d(e_i^f) = +1, d(e_{i+1}^f) = -1; \\ \mathbf{I}(s_r, s_{r'}) = 1 & \text{if} \quad d(e_i^f) = -1, d(e_{i+1}^f) = +1; \\ \mathbf{I}(s_r, t_{r'}) = 1 & \text{if} \quad d(e_i^f) = -1, d(e_{i+1}^f) = -1. \end{cases} \tag{2}$$

Assume all traces contain a total of $\mathcal{N}^a$ segments. That means there are $2\mathcal{N}^a$ vertices and the indicator matrix $\mathbf{I}$ is a $2\mathcal{N}^a \times 2\mathcal{N}^a$. As described in **Algorithm 4**, we examine the $2\mathcal{N}^a$ vertices one by one. If a vertex i has not yet been labelled, we label it as well as all other vertices with $\mathbf{I}(i, j) = 1$ with a new vertex ID. The result is a set of unique vertices $\mathbf{U} = \{U_1, ..., U_{\mathcal{N}^u}\}$. Knowing the set of unique vertices, we can further verify an unique edge as the edge connecting two distinct unique vertices. The set of unique edges is $\mathbf{E} = \{E_1, ..., E_{\mathcal{N}^e}\}$.

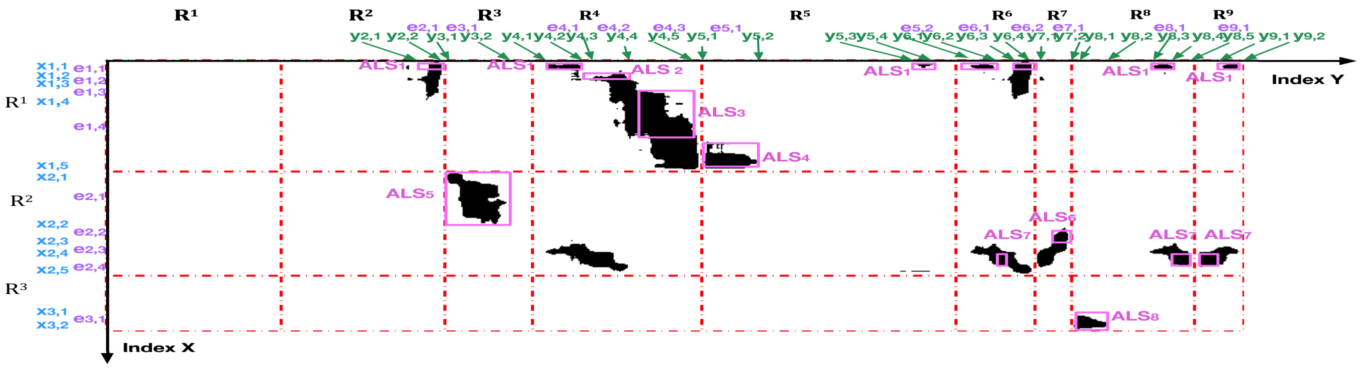
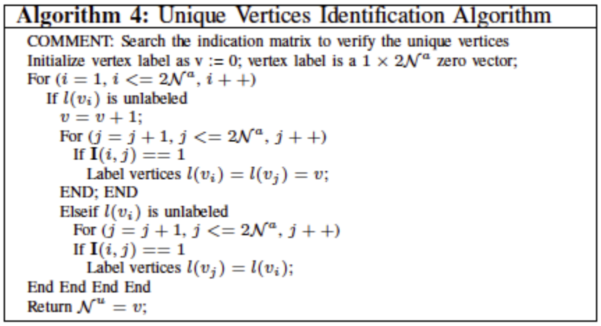


Figure 6. Relationship of BPs of HCPs and Trace Segments



Algorithm 4: Unique Vertices Identification Algorithm

```
COMMENT: Search the indication matrix to verify the unique vertices
Initialize vertex label as v := 0; vertex label is a $1 \times 2\mathcal{N}^a$ zero vector;
For ($i = 1, i <= 2\mathcal{N}^a, i++$)
  If $l(v_i)$ is unlabeled
    $v = v + 1;$
    For ($j = j + 1, j <= 2\mathcal{N}^a, j++$)
      If $\mathbf{I}(i, j) == 1$
        Label vertices $l(v_i) = l(v_j) = v;$
    END; END
    Elseif $l(v_i)$ is unlabeled
      For ($j = j + 1, j <= 2\mathcal{N}^a, j++$)
      If $\mathbf{I}(i, j) == 1$
        Label vertices $l(v_j) = l(v_i);$
End End End End
Return $\mathcal{N}^u = v;$
```

Figure 9. Unique Vertices Identification Algorithm

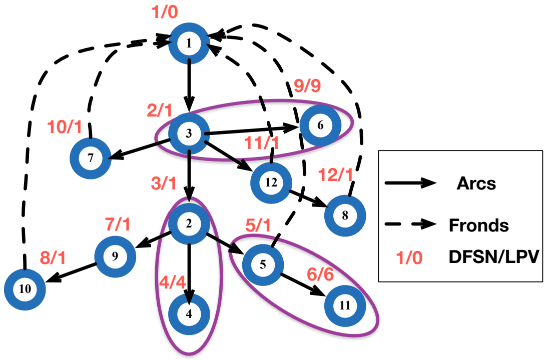


Figure 10. A sample of spanning tree

IV. Floor Map Construction/Visualization

With the set of vertices $\mathbf{U}$ and the set of unique edges $\mathbf{E}$, we create graph $\mathbf{G} = (\mathbf{E}, \mathbf{U})$. The graph is described by an adjacency matrix $\mathbf{A}$. The draft floor map is an embedding of $\mathbf{G}$ which is drawn according to the adjacency matrix.

The floor map construction algorithm aims to embed the graph $\mathbf{G}$ on a plane in a way that is visually more intuitive to a human observer. Three steps are involved: Depth First Block Search (DFBS), path finding, and straight-line embedding.

A. Depth First Block Search

This step is based on Tarjan's DFS block search algorithm [10]. The purpose is to label each vertex with a DFS number $DFSN(v)$, create a spanning tree, and identify blocks and fronds in the graph in order to enable path finding. DFS starts from the vertex with the highest node degree, and iteratively searches for unexplored descendants. Each new vertex is numbered by a DFS number according to the order in which it is explored. DFS stops when all the edges are explored. As result, the edges are separated into a set of arcs making up a spanning tree $\mathbf{T} = v \rightarrow w$, where $DFSN(v) < DFSN(w)$, and a set of fronds $\mathbf{F} = \mathbf{E} - \mathbf{T} = v \dashrightarrow w$, where $DFSN(v) > DFSN(w)$. In addition, DFS assigns an important parameter called the *low point value* to each vertex, and identifies the "blocks", which are the biconnected components of the graph. For vertex v, its *low point value* is defined as $LPV(v) = min(\{DFSN(v)\} \cup \{LPV(w)|v \rightarrow w\} \cup \{DFSN(w)|v \dashrightarrow w\})$ where initially all low point values are set to be the corresponding DFS number: $LPV(v) = DFSN(v)$. After the low point values are calculated, we look for all vertices such that $DFSN(v) \leqslant LPV(v)$. The edge leading to such a vertex v is a bridge, which is an edge whose deletion would partition the graph. The bridge and all edges whose connectivity to the graph depends on this bridge are grouped into a sub-graph called a block. All remaining edges are also grouped into a block. An example of a spanning tree is shown in Fig. 10. The numbers next to each node are the nodes DFS number and low point value respectively. There are four blocks in the example, and the largest block is the set of edges that exclude edges $(3, 6)$, $(2, 4)$, $(5, 11)$.

B. Path-Finding Algorithm

This algorithm searches for circle paths block-by-block, starting from the largest block. Initially, all vertices and edges in the block are marked as unexplored. We start from the vertex with the smallest DFS number in the block, which is the root vertex in the subgraph and mark it as explored. In each iteration, we extend the path to a neighbor vertex with unexplored edges. If the neighbor contains an unexplored frond, the path is outputted as a circle path. We repeat until all vertexes and edges in the block are explored. If there is only one edge which is in T in the block, we also output the edge as a path.

C. Planar Embedding Algorithm

This algorithm is based on a straight-line planar drawing algorithm [11]. Two main constraints are considered: i) Each edge must be a straight-line; ii) The angle between edges must have good angle resolution. The path-finding algorithm produces a set of distinct circle paths. The direction of the paths is ignored. We first draw the circle paths with four edges, followed by those with three and then five or more edges. Finally, the non-circle paths are drawn. The final result is a draft floor map, which would enable us to visualize the logical relationship of the vertices and edges on a plane.

V. Experiment and Results

The experiment took place at the lab area on the third floor of our academic building. The actual floor map of the area is shown in Fig. 2. The total survey area is 93 meters by 21 meters. In the experiment, a student is equipped with SAMSUNG GALAXY Tab 2 (7.0 version) and walks at a relatively constant speed around the area. A total of 20 traces are recorded with a total of 1468 measurements containing RSS values from 267 APs.

Following the framework of IMMS (shown in Fig. 1), the dimension of the measurement is reduced to 28. The traces segmentation and locations recognition algorithm identifies 38 ALSs in total. Then, the unique vertices identification algorithm produces 11 unique vertices and 13 unique edges.

Fig. 11 is the resulting draft floor map, and Table II shows the relationship of unique edges and ALSs. The two numbers next to the edge index are the number of times the edge appears in different traces and the total number of measurements corresponding to the edge. The black solid arrows are edges that appear more frequently and they match the corridors in the physical floor map quite well. The dash arrows are noise vertices and edges, which do not match the physical floor map. They apparently arise because of variability in the RSS signals of the APs. In the future, we may need to conduct more extensive experiments to determine how we may eliminate the noise vertices and edges or how we may merge them with those that match the physical map.

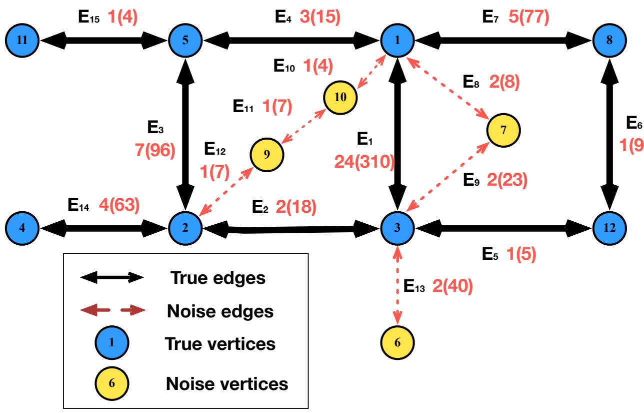


Figure 11. The resulting graphical floor map

VI. Conclusion and Future Work

In this paper, an automatic floor mapping system, IMMS, for draft floor map construction was presented. IMMS uses unlabeled crowd-sourced RSS measurements to construct the floor map of a building. Unlike existing fingerprinting methods, no elaborate manual off-line data collection process at fixed location is required. Accelerometers or other MEMS devices for measuring heading directions and distances are also not used. IMMS is an unsupervised system and is time efficient. The frequently found ALSs can be correctly correlated to corridor segments in the physical environment. Some noise vertices and edges are produced because of variability in the RSS signals. We need to conduct more extensive experiments and to enhance our algorithms so that these noise vertices and edges can be eliminated or merged with other ones.

TABLE II. Adjacency list of unique vertices

Vertex Inx.	Incident Edges	ALS Inx.
Node U_1	$\mathbf{E_1}$	1, 3, 22, 24
	$\mathbf{E_4}$	21
	$\mathbf{E_8}$	4
	$\mathbf{E_7}$	6
	$\mathbf{E_{10}}$	14
Node U_2	$\mathbf{E_2}$	17
	$\mathbf{E_{14}}$	10
	$\mathbf{E_3}$	11, 18
	$\mathbf{E_{12}}$	12
Node U_3	$\mathbf{E_{13}}$	29
	$\mathbf{E_9}$	2
	$\mathbf{E_5}$	24
Node U_5	$\mathbf{E_{15}}$	32
Node U_8	$\mathbf{E_6}$	35
Node U_9	$\mathbf{E_{11}}$	13

The next step of our work is to construct a radio map on top of the draft floor map. The radio map can then be used in on-line localization and tracking application of individuals.

References

[1] Y. Kong, Z. Zhong, G. Yang, X. Luo, A. K. S. Wong, and H. Zhai, “A non-parametric kernel method for CDMA2000 network indoor localization using multiple observations,” in Proc. Int. Conf. Inform Netwrking., 2012, pp. 97–101.

[2] R. W. Ouyang, A. K. Wong, C. T. Lea, and M. Chiang, “Indoor location estimation with reduced calibration exploiting unlabeled data via hybrid generative discriminative learning,” IEEE Trans. Mobile Comput., vol. 11, Sep. 2011, pp. 1613–1626.

[3] V. Y. Zhang, A. K. Wong, and K. T. Woo, “Histogram based particle filtering with online adaptation for indoor tracking in WLANs,” Int. J. Wireless Inform. Networks, vol. 19, no. 3, 2012, pp. 239–253.

[4] M. Panzarino. What exactly Wi-Fi SLAM is, and why apple acquired it. URL: http://thenextweb.com/apple/2013/03/26/what-exactly-wifislam-is-and-why-apple-acquired-it/#!p5wMH [accessed: 2014-05-20]. (Mar. 2013)

[5] H. Shin, Y. Chon, and H. Cha, “Unsupervised construction of an indoor floor plan using a smartphone,” IEEE Trans. Systems, Man, and Cybernetics-Part C: Applications andReviews, vol. 42, no. 6, Nov. 2012, pp. 889–898.

[6] M. Zhou, A. K. S. Wong, Z. Tian, Y. Zhang, X. Yu, and X. Luo, “Adaptive mobility mapping for people tracking using unlabelled Wi-Fi shotgun reads,” IEEE Commun. Lett., vol. 17, no. 1, 2013, pp. 87–90.

[7] R. W. T. Ouyang, A. K. S. Wong, M. Chiang, K. T. Woo, V. Y. Zhang, H. S. Kim, and X. M. Xiao, “Energy efficient assisted GPS measurement and path reconstruction for people tracking,” in Proc. IEEE GLOBECOM, 2010, pp. 1–5.

[8] E. Alpaydin, Introduction to Machine Learning, 2nd ed. Cambridge, Massachusetts, London, England: The MIT Press, 2010.

[9] K. Pearson, “Notes on regression and inheritance in the case of two parents,” in the Royal Society of London, vol. 58, Jun. 1895, p. 240?242.

[10] R. Tarjan, “Depth-first search and linear graph algorithms,” SIAM J. Comput., vol. 1, Jun. 1972, pp. 146–160.

[11] P. Rosenstiehl and R. E. Tarjan, “Rectilinear planar layouts and bipolar orientations of planar graphs,” Discrete Comput. Geom, vol. 1, 1986, pp. 343–353.

Evolving Future Internet Clean-Slate Entity Title Architecture with Quality-Oriented Control Plane Extensions

Jos Castillo Lema,
Felipe Silva
and Augusto Neto

Federal University of
Rio Grande do Norte (UFRN)
Natal, Brazil
Email: jcastillo@ppgsc.ufrn.br

Flavio de Oliveira Silva
and Pedro Frosi

Federal University of Uberlandia (UFU)
Uberlandia, Brazil
Email: flavio@facom.ufu.br

Carlos Guimares,
Daniel Corujo
and Rui Aguiar

Telecommunications Institute (IT)
Aveiro, Portugal
Email: cguimaraes@av.it.pt

***Abstract*—Due to the technological evolution, growth and various new service demands requiring new solutions to support novel usage scenarios, current Internet has been confronted with new requirements in terms of network mobility, quality and scalability, among others. New Future Internet approaches targeting Information Centric Networking, such as the Entity Title Architecture (ETArch), provide new services and optimizations for these scenarios, using novel mechanisms leveraging the Software Defined Networking (SDN) concept. However, the current ETArch approach is equivalent to the best-effort capability of current Internet, which limits achieving reliable communications. In this work, we evolved ETArch with both quality-oriented mobility and resilience functions following the super-dimensioning paradigm to achieve advanced network resource allocation integrated with OpenFlow. The resulting framework, called Support of Mobile Sessions with High Transport Network Resource Demand (SMART), allows the network to semantically define the quality requirements of each session to drive network *Quality of Service* control seeking to keep best *Quality of Experience*. The results of the preliminary performance evaluation of SMART were analyzed using Mininet, showing that it allowed the support of mobile multimedia applications with high transport network resource and quality demand over time, as well as efficiently dealing with both mobility and resilience events.**

Keywords–Future Internet; SDN; ICN; QoS and QoE.

I. Introduction

The Internet is constantly evolving, motivated by its natural growth and by the introduction of new services and applications to fulfill emerging needs. New requirements are being placed over its architecture, such as mobility, security and scalable content distribution. To cope with this new set of requirements, several enhancements are being defined, increasing the complexity of the overall Internet architecture, with many core components reaching their limit, and hindering further evolutions [1]. In addition, the current Internet still cannot address many of today's and emerging requirements adequately, such as efficient transmission of content-oriented traffic and effective congestion control. As a result, clean-slate attempts are being carried out as the next step towards an efficient Future Internet approach.

Information Centric Networking (ICN) [2] is one of such proposed approaches focusing on content access and delivery beyond current host-to-host communications. Content has a more central role in the network operations, motivated by the need to meet data-intensive applications. This paradigm shift leverages in-networking caching and replication, improving efficiency, scalability and robustness.

However deploying ICN capable nodes into current networks would require the update or replacement of existing networking equipment and protocols. *Software Defined Networking* (SDN) [3] emerges as a promising solution to overcome this, since it could not only facilitate the deployment of ICN functionalities in current networks without requiring new clean-slate designs, but it could also improve and enhance current and future Internet network management mechanisms.

The *Entity Title Architecture* (ETArch) [4] is an emerging Future Internet clean-slate approach which shares the vision of content-oriented paradigms, where entities request content by subscribing to it, triggering the network to dynamically configure itself in order to provide the users with the intended content. The content is delivered trough a channel that gathers multiple communication entities, called *Workspace*, allowing communicating entities to express their requirements over time. Despite its innovative approach, ETArch does not consider reliable communications provisioning in its design, and omits important factors to determine the connection, such as the quality requirements of demanding applications and the level of quality of the network nodes. Thus, ETArch lacks quality-oriented mechanisms for establishing workspaces, which means that network control functions seriously restrict data dissemination over the best-effort transport model of the current Internet. Moreover, ETArch operates in a per-flow driven way, and it is well known that such signaling approach overloads the system performance with the increasing session-flow admissions, mainly in terms of signaling and processing overheads [5]. As a result, the entire system can reveal increasingly high latency (network processing) and bandwidth use (exceeding signaling), which may increase energy consumption levels while degrading users perception.

This way, it is evident that ETArch is unable to accommodate bandwidth-intensive mobile session flows (e.g., real-time multimedia) guaranteeing both *Quality of Service* (QoS) and *Quality of Experience* (QoE) over time, in terms of setting workspaces connections with limited delay, error and loss rates experience. This drawback seriously restricts the scope of ETArch in Future Internet scenarios, especially when is taken into account the fact that traffic forecasts predict that 80% of the total data flows will stream multimedia content by 2017 [6]. In view of this, the session setup control functions of ETArch must take into consideration quality parameters to guide quality-oriented sessions, specially real-time ones, where losses above 5% generally lead to very poor effective throughput [7]. This diversity of applications makes the current ETArch approach of offering the same "best-effort" service to all applications inadequate.

The limitations described above motivate our work in the sense that there is a need to extend the control plane of legacy ETArch with quality-oriented functions to improve the session admission mechanism. First of all, it is required to define the application session requirements that will semantically describe the quality demands that must be fulfilled over time, by defining the minimum quality requirements of each mobile session flow (bitrate, tolerance to packet delay/loss/error, etc.). We claim that adopting both QoS-connectivity over-estimated provisioning capabilities and QoS-oriented mobility would benefit ETArch system to establish personalized multiparty sessions while improving the system scalability. For this reason, this paper proposes a new network architecture, denoted as Support of Mobile Sessions with High Transport Network Resource Demand (SMART), which redesigns the legacy ETArch with advanced QoS and mobility control functions to accommodate bandwidth-intensive mobile sessions over truly reliable and robust communication channels, while optimizing the network control plane. The SMART approach will act as a communication service provider with the following main innovations: (i) clean-slate Future Internet network architecture with new addressing methods, group-based connectivity, QoS-oriented mobility and resilience controls; (ii) IEEE 802.21 compliant signaling approach to control device handover; (iii) over-provisioning paradigm based automated, systematic and dynamic network resource allocation integrated with OpenFlow; (iv) OpenFlow extensions to provide QoS support.

The results of the preliminary performance evaluation of SMART were analyzed using Mininet, demonstrating its superior benefits with regard to the original configuration in the network and user perspectives in terms of QoE and delay.

The remainder of the document is organized as follows: Section II presents the background for this work, highlighting not only the supporting technologies, but also other related approaches. Section III presents the proposed framework, evaluated in Section IV, where results of its implementation are presented. Finally, Section V presents some concluding remarks.

II. Background

The Entity Title Architecture [4] is a clean-slate network architecture, distinguished over other Future Internet initiatives by its topology-independent naming, addressing and semantically driven designation scheme, that uniquely identifies each entity, and by the definition of a channel that gathers multiple communication entities, called *Workspace*. A key component of this architecture is the *Domain Title Service* (DTS), which deals with all network control-plane aspects. The DTS is composed of *Domain Title Service Agents* (DTSAs), which maintain information about entities registered in the domain and the workspaces that they are subscribed to, aiming to configure the network devices to implement the workspaces and to allow data to reach every subscribed entity.

The operation of ETArch, on which the DTSA entity centrally controls the behavior of the forwarding plane, materializes the SDN paradigm through OpenFlow. OpenFlow [8] is an instantiation of SDN already available in a number of commercial products and used in several Future Internet research projects. It separates the data plane from the control plane of the network, allowing the OpenFlow Controller to manage and control the underlying data plane, and to configure the forwarding table of the switches, via a well-known service-oriented API. This approach enables switches to be (re)configured on the fly, enabling flexible and dynamic network management [3].

The adoption of OpenFlow is mainly focused on core/wired networks. However, the support of QoS in OpenFlow-enabled networks is very limited, relaying on manual external tools to manage queue configuration. Several recent attempts have tried to overcome such limitation, such as QoSFlow [9], that made possible for administrators to manage resources on the controller level.

A. Related work

Regarding QoS, there is a continuing debate on how to evolve the current Internet in order to efficiently accommodate multimedia sessions. Currently, there is no QoS architecture that is successful and globally implemented. Some researchers argue that fundamental changes should be done to fully guarantee QoS, while others think slight changes are enough to have soft guarantees which will provide the requested QoS with high probability. Future Internet requires QoS control approaches beyond current Internet standards, which mainly leverage the per-flow approach to allocate network resources (queues, bandwidth, data paths, etc.). Drawbacks associated to per-flow approaches are well known [5], mainly in terms of network performance (state, processing and signaling overheads), severely jeopardizing system scalability and increasing energy consumption.

Our previous works [10] proposed dynamic super-dimensioned provisioning network resource allocation techniques, deploying a controlled oversizing strategy for both bandwidth and data paths and allowing the admission of several sessions without per-flow signaling exchanges and decisions in the entire network systems. We strongly believe that an optimized network control approach enabled by the over-provisioning technique will allow the evolution of ETArch towards a truly efficient and robust Future Internet network system in comparison to what it is available in the literature.

Several works have explored QoS control and OpenFlow integration in Future Internet architectures, as follows. B. Sonkoly et al. [11] focus on enhancing OpenFlow switches and OpenFlow testbeds with advanced QoS and virtualization capabilities, in order to make them capable of running QoS related experiments, but does not propose any specific QoS control model. In the other hand, H. Egilmez et al. [12] propose a per-flow driven approach, while our focus is to conceive QoS control mechanisms beyond IP and per-flow regular approaches.

In [13], key research topics in the area of future Internet architecture are investigated. The most relevant research projects from United States, European Union, Japan, China, and other countries are introduced and discussed, aiming to draw an overall picture of the current research progress on the Future Internet architecture. Among all of them, only the Japanese proposal AKARI briefly mentions QoS in the design principles of one of its sub-architectures. Not only clean-slate proposals are not focusing on QoS (neither QoE), but most of them are not even taking it under consideration.

The analysis of the related work justifies our work, since none of the proposals taken into consideration fulfills the requirements in providing a Future Internet clean-slate SDN system supporting truly reliable and robust bandwidth-intensive transport capacities.

III. SMART Proposal

The SMART has as main objective to enhance ETArch with new mechanisms supporting advanced network control capabilities aiming to enable QoS-guaranteed mobile multimedia applications over time. Quality requirements are semantically defined for each session in order to guide SMART functionalities, supported by an extended OpenFlow approach to support QoS control.

The SMART envisions enabling a new integrated Future Internet clean-slate SDN system embedding new mechanisms to support advanced routing, resource reservation, admission control and priority queuing functionalities. In order to fulfill the required end-to-end QoS, we designed a dynamic QoS routing super-dimensioned provisioning centric strategy to provision automated, systematic and dynamic network resource allocation for multimedia workspaces.

The innovating aspect of the advanced QoS control adopted in SMART focuses on enabling the integrated use of admission control and over-provisioning centric network resource allocation to achieve a signaling constrained approach. SMART bootstraps the system with oversized network resources, namely surplus workspaces enforced with over-reservations on all network interfaces, and stores such information in the DTSA. As such information is available in advance, the DTSA is enabled to take multiple session admission decisions without any signaling events to enforce neither resource reservations nor forwarding rules in the selected workspace. After the system bootstrap (at the network boot up), SMART only generates signaling events to adjust the over-reservation patterns, in order to over-provision the system again, allowing multiple session admissions with the least amount of signaling.

The SMART framework is presented in Figure 1, emphasizing the new QoS-Manager, which embeds the QoS control-plane additions.

The DTSA acts as the OpenFlow controller of the network. In what concerns its functions as OpenFlow controller, the DTSA is responsible for storing information about the existing entities (Entity Manager), workspaces (Workspace Manager) and handover procedures (Mobility Manager), as well as for performing routing related tasks, implementing the workspaces into the switches. Moreover, these functions are interfaced by a central module (NetConnector), allowing the integration of procedures to optimize several aspects of the network. Lastly, it features a *Media Independent Handover Function* (MIHF) for exchanging IEEE 802.21 information with other nodes and an OpenFlow Channel for communication with the OpenFlow Switches. The IEEE 802.21 is the IEEE standard for Media Independent Handover (MIH) [14]. Its main purpose is to facilitate and optimize inter-technology handover processes by providing a set of media-independent primitives for obtaining link information and controlling link behavior in a heterogeneous way, thus creating an abstraction regarding the link layer.

The EDOBRA Switch consists of an IEEE 802.21-enabled OpenFlow switch. Besides the standard OpenFlow switch capabilities for executing data packet for- warding operations and for storing information on how packets of each workspace should be treated, the EDOBRA Switch is coupled with IEEE 802.21 mechanisms to control aspects of the link interface regarding handover management, such as resource management and/or events about the attachment and detachment of nodes. Lastly, it is coupled with an MIHF for interacting with the Mobile Node (MN) and the DTSA via IEEE 802.21 and an OpenFlow Channel for communication via OpenFlow with the DTSA. The OpenFlow Channel is also responsible for encapsulating DTS messages into OpenFlow messages.

The Mobile Node represents the end-user equipment that establishes connection with the endpoint switches. The MN may be equipped with one or more access technologies, either wired (e.g., Ethernet) or wireless (e.g., WLAN or 3G). The MN deploys an MIHF, allowing higher-layer entities in the device itself (Mobility Manager) or external network entities (e.g., DTSA) to control the links and to retrieve information in an abstract way. In this way, the MN is able to either retrieve link conditions on the current connection or to provide information about other networks in its range. In what concerns DTS procedures (such as register, workspace creation and attachment operations), the MN contains a DTS Enabler that allows it to communicate with endpoint switches via DTS. In addition, the DTS Enabler is also used by applications to send their packets over DTS protocol.

The proposed new sub-components of the QoS-Manager are described as follows:

Advanced Resource Allocator: The QoS Advanced Resource Allocator provides support to the QoS management by controlling the usage of the network resources. It is responsible for calculating the new over-reservation patterns, in case none of the available workspaces can possibly satisfy the QoS requirements of a new session; and for the enforcement of the new over-reservation patterns over the workspace switches through the Protocol Manager.

Admission Controller: The QoS Admission Controller provides support to the network's QoS management by regulating the access to the network. It is responsible for querying session requirements, candidate paths and their resource availability; and for taking the final decision, either accepting or rejecting the establishment of the workspace. The minimum quality requirements for each mobile session flow (bitrate, tolerance to packet delay/loss/error, etc.) and the current conditions of the candidates workspaces (available traffic classes,

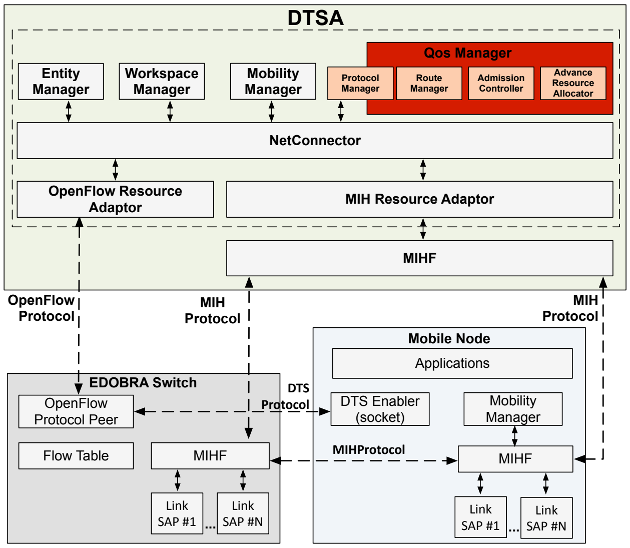


Figure 1: Proposed framework

packet delay/loss/error current rates, link technology, etc.) are taken into account.

The Admission Controller denies a session when the demanded QoS parameters cannot be satisfied (i.e., there is no feasible workspace to accommodate the session), and informs the controller to take necessary actions.

Route Manager: This function is responsible for determining the availability and packet forwarding performance of routers to aid the route calculation. It requires collecting the up-to-date network state from the switches on a synchronous or asynchronous basis. Several routing algorithms, such as shortest path or a dynamic QoS-aware one, can run in parallel to meet the performance requirements and the objectives of different sessions. Network topology information is needed as input along with the service reservations.

Protocol Manager: The QoS Protocol Manager handles communication between the QoS-Manager and the extended OpenFlow API. It is responsible for the setup of the over-estimated reservation patterns across the network through the extended OpenFlow API, for collecting the flow definitions received from the QoS-Manager and for efficient flow management by aggregation.

SMART was designed under the principle of pushing complexity to the network boundary (application hosts, leaf or first-hop routers and edge routers). Since a network boundary has a relatively small number of flows, it can perform operations at a fine granularity, such as complex packet classification and traffic conditioning. In contrast, a network core router may have a larger number of flows, it should perform fast and simple operations. The differentiation of network boundary and core routers was accomplished through workspace aggregation, and it is vital for the scalability of SMART.

A. System Setup

The System Setup is triggered by the DTSA as a consequence of noticing that the underlying network topology has changed. Therefore, DTSA agents unicast an OpenFlow extended message to all OpenFlow enabled switches in the network. On receiving the message, each switch initializes the per-class over-estimated reservation patterns in a way compatible with the underlying QoS approach (for instance, configuring the packet scheduling priorities).

At this stage, the DTSA polls each switch of the network. The current condition of each switch must be taken into account (available traffic classes, packet delay/loss/error current rates, link technology, ect). When the stats request is responded, the DTSA stores all the information in local state tables (unicast workspaces at this time). The generation of multicast workspaces is still a part of the System Setup, which is a fundamental support for the workspace selection. To that, DTSA adopts a combinational algorithm that takes unicast workpace registers to generate all possible combinations between each ingress and all core/egress sequentially.

B. Session Setup

This process is triggered whenever the DTSA receives a workspace attachment entity request.

It is necessary to decide the best-suited path in the core network in order to maintain the established QoS parameters of the multiparty content delivery (as described in Figure 2). An efficient approach to quality-oriented mobility control must always keep the mobile nodes best connected over time, and guarantee that the whole activated mobile session flow meets its quality requirements. The algorithm starts by searching in the internal structures of the DTSA to determine whether there is already a workspace that is being used for the specified flow from the traffic source to the subscriber.

```
Query QoS requirements of the entity attachment request;
Get all workspaces from the traffic source to the subscriber;
for each candidate workspace do
    if workspace able to acommodate QoS reqs then
        Configure workspace flow using OpenFlow in source switch;
        Consigure workspace flow using OpenFlow in dest switch;
        Join the user to the existing workspace;
        break;
for each candidate workspace do
    if readjusted workspace able to acommodate QoS reqs then
        for each switch needed of readjustment do
            Setup new over-reservation patterns(extended OpenFlow)
        Configure workspace flow using OpenFlow in source switch;
        Consigure workspace flow using OpenFlow in dest switch;
        Join the user to the existing workspace;
        break;
Reject the entity attachment request;
```

Figure 2: Session setup algorithm

If there is indeed a workspace able to acommodate the QoS requirements of the new session-flow, it is only necessary to join the user to the existing workspace, which requires no significant signalization overhead (only end switches are notified), as opposed to the original ETArch architecture without the SMART extensions, in which all switches forming the workspace must be signalized.

Considering the case of non-existence of available workspaces to accommodate the demanded session, a suitable workspace may be found by simply readjusting the current over-reservation patterns. The workspace with a greater probability of acceptance is selected and the new over-reservation configuration is calculated, as can be seen in (1).

$$B_{ov}(i) = \frac{B_u(i)}{MR_{th}(i)}(MR_{th}(i) - B_u - B_{rq}(i)) \quad (1)$$

$$\begin{aligned} where\ B_{ov} &: Overreservation\ Bandwidth\ of\ CoS\ i; \\ B_u(i) &: Bandwidth\ Used\ in\ CoS\ i; \\ B_{rq}(i) &: Bandwidth\ Required\ in\ CoS\ i; \\ MR_{th}(i) &: Maximum\ Reservation\ Threshold\ of\ CoS\ i \end{aligned}$$

If the over-reservation patterns calculated by the DTSA are not enough to ensure a suitable path, it is necessary to make a readjustment of the maximum reservation thresholds of all the classes. When none of the available paths are able to accommodate the demanded session (not enough bandwidth in the network) and there is no workspace candidate with probability of acceptance, DTSA rejects the entity attachment request.

IV. Evaluation

In order to evaluate the feasibility of our framework, we extended the ETArch implementation with the SMART architecture according to the proposals in Section III.

A. Evaluation Scenario

The results of the preliminary performance evaluation of SMART were analyzed by using Mininet [15]. As presented in Figure 3, two different *Mobile Nodes* (MN) were connected to a common OpenFlow Switch.

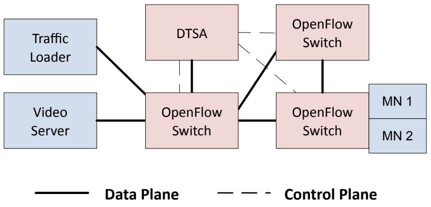


Figure 3: Scenario environment description

The DTSA is connected to the OpenFlow devices using two different connections: one for control and another for data. The MN1, MN2 and the Video Server, on which the DTS applications were run, are the remaining entities that complete the evaluation scenario. The application in the video server is sending a H.264 video stream over two workspaces, one of them being a QoS-enabled workspace, with the MN1 and the MN2 subscribed to each of them in order to receive the video stream. The switches are connected in a triangular shape to have path diversity. The video streaming server and the client are connected to different switches, while the traffic loader inserting cross-traffic into the network is connected to the same switch that the server connects to. Each switch initiates a secure connection to the controller using the OpenFlow protocol (see dashed lines in Figure 3). The controller runs our SMART implementation described in detail in Section III.

In this scenario, MN1 requests a QoS-enabled workspace to receive content from the Video Server, while MN2 requests a normal workspace with no special QoS requirements. Thus, the video packets destined to MN1 are identified as being part of a multimedia workspace by the SMART controller and routed accordingly, while the stream (destined to MN2) is considered as a data workspace which has no QoS support (i.e., best-effort). Finally, in each test, long cross-traffic is sent from the loader to the client continuously.

B. Performance Evaluation

In this section, we evaluate the performance of the proposed framework, comparing it with a deployment of the ETArch without QoS support. Throughout the tests, we used a video sequence having 30 frames per second with the resolution of 1280x720. We then encoded the sequence in H.264 format using the *ffmpeg* encoder (v.1.2.4) to obtain a stream at 1800 kbps (32.55dB).

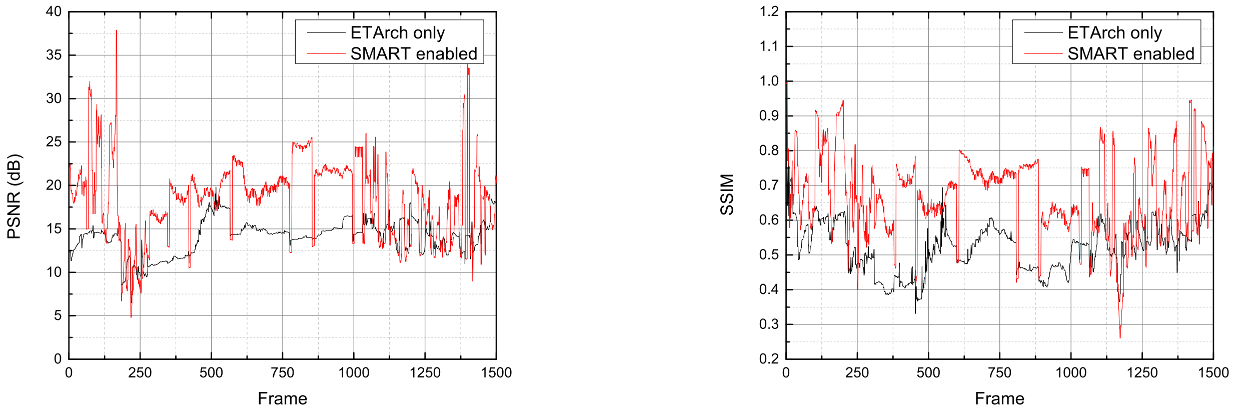
We decoded the received videos using *ffmpeg* and measured their qualities using *Peak Signal-to-Noise Ratio* (PSNR) and *Structural Similarity* (SSIM) values with respect to the original raw video. PSNR is the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. It is widely used to measure the quality of reconstructed transmitted images/videos. The SSIM index is a method for measuring the similarity between two images. It is a full reference metric; in other words, the measuring of image quality based on an initial uncompressed or distortion-free image as reference. SSIM is designed to improve on traditional methods like PSNR and *Mean Squared Error* (MSE), which have proven to be inconsistent with human eye perception. The results are given in Figure 4, which are in terms of received video quality versus time.

Results show that the video with QoS support (SMART enabled) is not affected from the cross traffic and approaches full video quality, while the video without QoS support (ETArch only) has a significant amount of quality loss. In terms of PSNR (Figure 4(a)), the original ETArch framework achieved 19.02±9.03 dB, while the SMART-enabled version achieved 20.97±7.94 dB. SMART achieved optimized bandwidth-guaranteed multimedia transport with a 10% of PSNR improvement. In terms of SSIM (Figure 4(b)), ETArch achieved 0.61±0.16, while SMART-optimized version 0.79±0.14. This implies an improvement of almost 30%. These results were achieved because, during the bootstrapping procedure, switches composing candidate workspaces were initialized with per-class over-reservation patterns. Besides, the video packets destined to MN1 are identified as being part of a multimedia workspace by the SMART controller and routed accordingly, while the stream (destined to MN2) is considered as a data workspace which has no QoS support (i.e., best-effort).

Figure 5 shows random frames picked for both streams. Figure 5(a) corresponds to the traffic subscribed by MN2 (without QoS support), while Figure 5(b) corresponds to the SMART enabled transmission.

C. Forwarding Table Size Analysis

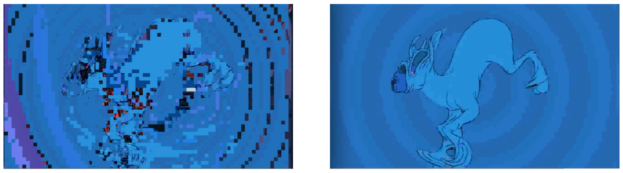
In this section, we study the footprint of the proposed framework, comparing it with a deployment of the ETArch



(a) PSNR metric

(b) SSIM metric

Figure 4: QoE metrics for multimedia video streaming



(a) Frame without SMART

(b) Same Frame using SMART

Figure 5: Video snapshot comparison

TABLE I: FORWARDING TABLE SIZES AT CORE ROUTERS

	ETArch with SMART	ETArch only
University of Texas (2011 report)	952	20.000
University of Texas (2013 report)	1.330	40.000
University of Texas (congested scenario)	1.330	80.000

without QoS support. As explained in Section III, SMART was designed under the principle of pushing complexity to the network boundary. Besides, the larger the size of the forwarding table, the worse the performance achieved. Studies show the performance degradation with the increasing number of flows [16].

The results obtained are presented in Table I, showing the number of entries for each protocol generated in the forwarding table for different scenarios. We use real-world scenarios, from different campus networks around the world. According to the 2011 report [17], the architecture of the campus network of the University of Texas consisted of 14 cores and 2 border routers, supporting up to 20.000 simultaneous connections. In 2013, the campus network architecture grew up to 16 cores and 2 border routers. Over 40,000 simultaneous connections spending 36 million hours combined on the system were monitored in spring of 2013. Let's also imagine a congestion scenario with twice as many simultaneous connections (80.000).

Results from Table I show a very significant optimization in the forwarding table size of core routers. However, the number of entries in the forwarding table of SMART signaling scheme does not depend on the number of entities attachments requirements, unlike the original ETArch signaling scheme. Therefore, the relative percentage of the forwarding tables size comparison could be even lower in more saturated scenarios. In what concerns the DTS protocol, no control signaling was required on core switches during the session setup procedure since over-estimated reservations patterns were already initialized during the bootstrapping procedure.

V. Conclusion

We have presented a QoS-enabled framework that aims to support mobile multimedia applications with guaranteed QoS and QoE on top of ETArch, a clean-slate SDN-based ICN approach. It allows the dynamic and preemptive reconfiguration of the network resources using over-estimated reservation patterns to achieve optimized bandwidth-guaranteed multimedia transport. Results showed that our framework allows mobile multimedia applications with guaranteed QoS maintained over time, optimizing traffic control and diminishing overhead and forwarding table sizes at core network switches. Moreover, using our framework, applications become semantically capable of defining the quality requirements of each session.

The work presented in this article showcased the integration and growth capabilities of multiple technologies, exposing them to novel scenarios, contribution to the evolution of SDN, ICN, mobility and QoS management procedures operating as a suitable Future Internet framework embodiment.

Acknowledgment

This work has been partially funded by the European Community's Seventh Framework Programme, under grant agreement n. 258365 (OFELIA project), and by Brazilian agencies CAPES and FAPEMIG.

REFERENCES

[1] M. Handley, “Why the internet only just works,” BT Technology Journal, vol. 24, no. 3, Jul. 2006, pp. 119–129.

[2] Information-Centric Networking Research Group (ICNRG). The internet engineering task force.

[3] M.-K. Shin, K.-H. Nam, and H.-J. Kim, “Software-defined networking (sdn): A reference architecture and open apis,” in ICT Convergence (ICTC), 2012 International Conference on, 2012, pp. 360–361.

[4] F. Silva, M. Goncalves, J. Souza, R. Pasquini, P. Rosa, and T. Kofuji, “On the analysis of multicast traffic over the entity title architecture,” in 2012 18th IEEE International Conference on Networks (ICON), 2012, pp. 30–35.

[5] J. Manner and X. Fu, “Analysis of existing quality-of-service signalling protocols,” RFC 4094, Internet Engineering Task Force, May 2005.

[6] Cisco, “Cisco visual networking index: Global mobile data traffic forecast update, 2012-2017,” 2013.

[7] J. Babiarz, K. Chan, and F. Baker, “Configuration guidelines for diffserv service classes,” RFC 4594, Internet Engineering Task Force, 2006.

[8] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, “Openflow: enabling innovation in campus networks,” SIGCOMM Comput. Commun. Rev., vol. 38, no. 2, Mar. 2008, pp. 69–74.

[9] A. Ishimori, F. Farias, E. Cerqueira, and A. Abelem, “Control of multiple packet schedulers for improving QoS on OpenFlow/SDN networking,” in 2013 Second European Workshop on Software Defined Networks, vol. 0. Los Alamitos, CA, USA: IEEE Computer Society, 2013, pp. 81–86.

[10] J. Castillo-Lema, E. Cruz, A. Neto, and E. Cerqueira, “Advanced resource provisioning in context-sensitive converged networks,” in 2013 International Conference on Computing, Networking and Communications (ICNC), Jan. 2013, pp. 77–81.

[11] B. Sonkoly, A. Gulyas, F. Nemeth, J. Czentye, K. Kurucz, B. Novak, and G. Vaszkun, “OpenFlow virtualization framework with advanced capabilities,” in 2012 European Workshop on Software Defined Networking (EWSDN), 2012, pp. 18–23.

[12] H. Egilmez, S. Dane, K. Bagci, and A. Tekalp, “OpenQoS: an OpenFlow controller design for multimedia delivery with end-to-end quality of service over software-defined networks,” in Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, 2012, pp. 1–8.

[13] J. Pan, S. Paul, and R. Jain, “A survey of the research on future internet architectures,” IEEE Communications Magazine, vol. 49, no. 7, 2011, pp. 26–36.

[14] LAN/MAN Committee of the IEEE Computer Society, “IEEE Std 802.21-2008, Standards for Local and Metropolitan Area - Part 21: Media Independent Handover Services,” 2008.

[15] B. Lantz, B. Heller, and N. McKeown, “A network in a laptop: Rapid prototyping for software-defined networks,” in Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks, ser. Hotnets-IX. New York, NY, USA: ACM, 2010, p. 19:119:6.

[16] A. Bianco, R. Birke, L. Giraudo, and M. Palacin, “OpenFlow switching: Data plane performance,” in 2010 IEEE International Conference on Communications (ICC), 2010, pp. 1–5.

[17] I. T. Services, “Campus network report. academic year 2010-2011,” University of Texas, Tech. Rep.

UML-based Modeling Entity Title Architecture (ETArch) Protocols

Diego Alves da Silva,
Natal Vieira de Souza Neto,
Flávio de Oliveira Silva
and Pedro Frosi Rosa

Faculty of Computing
Federal University of Uberlândia
Uberlândia, MG, Brazil
Email: diegoalves@cti.ufu.br,
natal@mestrado.ufu.br,
flavio@facom.ufu.br,
pfrosi@ufu.br

Michel dos Santos Soares

Faculty of Computing
Federal University of Sergipe
São Cristóvão, SE, Brazil
Email: mics.soares@gmail.com

***Abstract*—The approaches used to model communication protocols suffered several changes in past years. Some of the modeling languages are not used anymore because of their complexity, others because of their inherent limitations that were pointed out over time. Even today an approach that can represent a protocol in many abstraction levels is welcome. The objective of this article is to introduce an approach to model a communication protocol using the Unified Modeling Language (UML). The purpose is to create models that are able to represent an Internet architecture in many abstractions levels and different concerns, including structural level and services definitions. Besides, we propose an evaluation of the generated model, showing main advantages, such as representing architectural modeling, and limitations, such as representing time, non-functional constraints and physical resources when modeling communication protocols using UML.**

Keywords-UML; protocols; architectural modeling

I. Introduction

The design and development of a real-time communication protocols must ensure security, reliability and response time capability. In other words, the protocol must not reach unsafe or not allowed states, must forecast all possible states and must comply with time constraints. With the purpose of getting a high abstraction level of states and message exchanging between them, and also to allow the validation of the required properties, different modeling languages were considered for modeling real-time protocols in past years, i.e., State Machines [1][2], Petri Nets [3], and LOTOS [4].

The mentioned languages have the same modeling purpose, they are all focused on modeling behaviour, but with little focus on the structural and architectural elements, i.e., the Petri Nets language is fully visual; however it does not have natively ways of modeling time constraints and architectural representations. Then, in order to solve issues such as time constraints, customizations of the language as Time Petri Nets and Coloured Petri Nets [5] were created. Another modeling language, the Language of Temporal Ordering Specification (LOTOS), is tightly algebraic and has a specification with complex symbology, which may be one of the causes why the language is not adopted as default language for protocol modeling [6].

The International Organization for Standardization(ISO)/ Open Systems Interconnection (OSI) reference model for communication has several problems [7][8]. These issues are at the network layer, mainly related to the Transmission Control Protocol (TCP)/ Internet Protocol (IP) model. This model basically consists of five layers: application, transport, network, link and physical. Each layer is responsible for ensuring a portion of the service, and does not guarantee that it is sent to the backsheet to solve this. Much of the services are made in the application layer, but some could be made in layers over the network core, such as the transportation and network layers [9]. This paper presents the modeling of a Future Internet protocol architecture using the UML modeling language, as well as standards and an approach to model elements and behaviours in different abstraction levels.

The reminder of the paper is as follows. In Section 2, an overview of related works about protocol modeling is described. In Section 3, the architecture and the services of the proposed Etarch protocol are presented. In Section 4, the modeling of services provided by the protocol is described using the UML modeling language. In Section 5, the modeling is evaluated and the advantages and limitation are discussed.

II. Related Works

Finite State Machine (FSM) consists of a mathematical model of computing. The FSM have an alphabet of input and output, states, and transitions that connect states. With a finite number of states, its main feature is determinism. Modeling communication protocols using State Machines consists of dividing the system into communicating components, in which

ISBN: 978-1-61208-360-5

each component is a State Machine. One advantage of this approach is the possibility of automatic validation of the model. The main limitations are the low abstraction level and the problem of the high number of created states to represent operations between components. Wu and Loui [10] presented the idea of modeling asynchronous protocols for communication across unreliable channels using finite-state machines communicating via an unreliable shared memory. It is shown that there are robust protocols for deletion and insertion errors. The state machine and intermediate variables were applied to solve the problem of difficulty of regulate the input variables and complex properties that can not be described in temporal logic and verify the related properties of the data flow control module, overcoming the incompleteness of the traditional methods [11].

Another modeling language that has been applied to model protocols is the Timed Automata [12], which consists in an approach of State Machine to treat time and clock modeling. The UPPAAL environment allows modeling and validation of Timed Automata models [13][14]. One of main properties of Timed Automata is that, although the set of configurations is in general infinite, checking reachability properties is decidable. However, an animation of Timed Automata cannot be determined, and inclusion checking is undecidable [12], except for deterministic timed automata. This basically forbids the use of timed automata as a specification language [15].

Since its introduction in 1962, but mostly after 1985, Petri Nets, a graphical and mathematical language, has been widely used to model communication protocols [16]. The language provides interesting modeling possibilities for real-time communication protocols, such as directly supporting modeling of concurrency, resource sharing and asynchronous events. The absence of compositionality is the main criticism raised in models created using Petri Nets [17][18]. Therefore, the level of abstraction is relatively low when comparing with UML. In addition, ordinary Petri Nets are not able to model temporal constraints [19]. In order to deal with the time modeling limitation, time extensions were proposed to the basic Petri net theory. The modular modeling of real-time communication protocol can be made using Time Petri Nets [20]. Another example is the Time Petri Nets model with Register (TPNR), which allows modeling of communication time delay [19]. However, this approach has a limited time structure such as the representation of composition time.

LOTOS allows the creation of many ways of transformation and validation of communication protocols. There are some examples of services of protocols implemented in LOTOS [21][22]. All approaches of LOTOS share the same problem, namely, the complexity of models. Besides, as the model is created in the early phases of a project, this property may difficult the construction of a complete model [23].

The UML [24] is currently widely applied in the software industry [25]. There are several approaches that use UML models as base for protocols [26][27][28]. Furthermore, UML is an extensible language, which makes it possible to create stereotypes and data types using the language metamodel. The mechanisms of extensibility allow to customize and extend UML resources, adding new building blocks, properties and specifying a new semantic, turning the UML adequate to specific domains. The Logical Link Control and Adaptation Layer Protocol (L2CAP) for wireless channel with bluetooth technology was modelled using UML, more specifically using the Sequence and State diagrams [27].

As was previously described, many modeling languages were applied to model communication protocols, and each one with specific characteristics. There are languages with focus on definitions of algebraic expressions and others on behavior and states. Therefore, in order to model a real-time communication protocol, it is necessary to use a modeling language that is capable of representing a lower abstraction level of modeling, including algebraic expressions, a model structure, robust time transformation and an abstract model. However, from the user point of view, it is also necessary to use a modeling language that is capable of modeling higher levels of abstraction, which makes more sense to the end user. Most commonly used modeling languages for communicating protocols lacks these characteristics.

III. ETARCH ARCHITECTURE

The Entity Title Architecture [29] (ETArch) is a clean state network architecture, where naming and addressing schemes are based on a topology-independent designation that uniquely identifies an entity, called Title, and on the definition of a channel that gathers multiple communication entities, called Workspace. A key component of this architecture is the Domain Title Service (DTS), which deals with all control aspects of the network. The DTS is composed by Domain Title Service Agents (DTSAs), which maintain information about entities registered in the domain and the workspaces that they are subscribed to, aiming to configure the network devices to implement the workspaces and to allow data to reach every subscribed entity.

Through ETArch, communications are handled by the Workspace. Therefore, ETArch inherently allows the integrated support of multicast and mobility within the Workspace that can be viewed as a logical bus interconnecting multiple entity instances (e.g., a service, a sensor, a smartphone, a host, or even a process). Its behavior is inspired by the multicast technology, where data is sent once by a source to the workspace, and all associated entities will receive.

The operation of ETArch, on which a centralized entity is responsible for the behavior of the forwarding plane, meets Software-Defined Networking (SDN) concepts [30], implemented in ETArch by the OpenFlow. OpenFlow [31] is an instantiation of SDN already available in a number of commercial products and used in several research projects. It separates the data plane from the control plane of the network, allowing a separate entity (i.e., the OpenFlow Controller) to manage and control the underlying data plane, configuring the forwarding table of the switches, via a well-known service-oriented API. This enables switches to be (re)configured on the fly, enabling flexible and dynamic network management [32] and allowing to bring life to the workspace driven communication concept.

Considering the ETArch networking model, the network itself is composed by several DTSAs that are configured in the model tree. When a workspace is requested by an entity that does not have DTS and workspace, it prompts the DTS higher-level information from that workspace, and the DTS asks the next level and so on, in a structure similar to the Domain Name System (DNS) used nowadays [33].

In order to support its concepts, ETArch defines protocols in the data and control plane. In the control plane, the signaling approach provides the services related with the life cycle of entities and workspaces, such as to register an entity at the Domain Title Service (DTS) or to create a workspace, attach and detach entities to a given workspace.

The Entity Title Control Protocol (ETCP) is responsible for the communication between an entity and the Domain Title Service Agent (DTSA), while the DTS Control Protocol (DTSCP) is responsible for the communication between DTSAs inside the DTS.

A. *Main ETCP primitives*

- ENTITY_REGISTER: Registers an entity at the DTS. To be registered an entity must present its title, capabilities and communication requirements. To communicate the entity must first register itself.
- WORKSPACE_CREATE: Creates a workspace locally at the DTSA. If the workspace has a public access after the successful creation, DTSA will advertise the workspace by inserting an entry at the Workspace Database.
- WORKSPACE_ATTACH: Attaches an entity to a workspace. To accomplish the attachment process, the DTSA will obtain all network elements and will configure them to extend that workspace. If the DTSA does have the information about the workspace, using the DTSCP protocol, it will send a WORKSPACE_LOOKUP primitive.
- ENTITY_UNREGISTER: Removes an entity from the DTS.
- WORKSPACE_DETACH: Removes an entity from an existing workspace.
- WORKSPACE_DELETE: Deletes a workspace and performs all clean up necessary at the NE of the current DTSA.

B. *Main DTSCP primitives*

- WORKSPACE_LOOKUP: Sent by a DTSA to its resolvers, i.e., the other DTSAs
- WORKSPACE_ADVERTISE: Inserts, deletes, or updates the Workspace Database, by indicating that a DTSA is part of the DTSA set of a specific workspace. The Operation receives the level indicating the visibility of that workspace. The DTSA stored at the Workspace Database must be of the same level or can be a Master DTSA of the level right below.
- DTS_MESSAGE: Enables communication between different DTSAs inside the DTS. If the DTSA source knows the path to the DTSA destination, this path will be contained in the message header. Otherwise, the message will be forwarded to the resolvers, until one of them knows how to compute the path to the destination DTSA. If the Master DTSA of the Root Level cannot compute the path to destination, the message will fail.

IV. Case Study

The modeling of the Etarch architecture protocol using UML aims to present the structure of the elements, behaviours and time constraints in a high abstraction level. For this, in this paper, the Class, Sequence and Composite Structure diagrams are presented.

The Class Diagram is responsible to define a classifier. Within this diagram, it is possible to define attributes, methods, visibility and relationship between one and many classifiers. The Class diagram is important in protocol modelling to define the used types, the data structures and the defined elements, and how they relate to each other. This is also useful for modeling the behavior of the protocol.

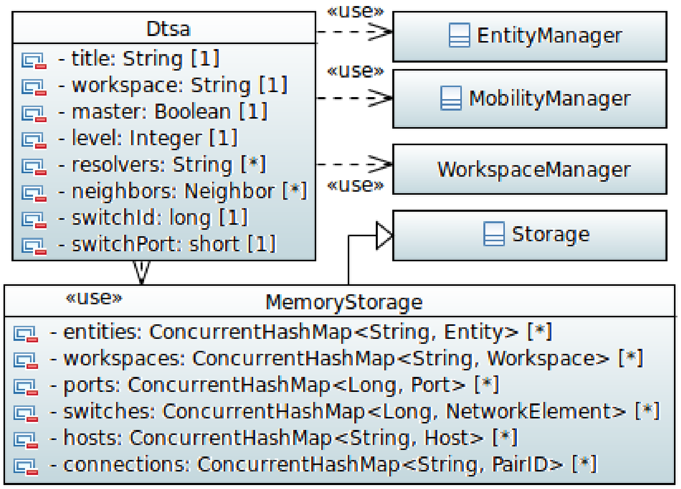

Dtsa
- title: String [1]
- workspace: String [1]
- master: Boolean [1]
- level: Integer [1]
- resolvers: String [*]
- neighbors: Neighbor [*]
- switchId: long [1]
- switchPort: short [1]
«use»
EntityManager
«use»
MobilityManager
WorkspaceManager
«use»
Storage
«use»
MemoryStorage
- entities: ConcurrentHashMap<String, Entity> [*]
- workspaces: ConcurrentHashMap<String, Workspace> [*]
- ports: ConcurrentHashMap<Long, Port> [*]
- switches: ConcurrentHashMap<Long, NetworkElement> [*]
- hosts: ConcurrentHashMap<String, Host> [*]
- connections: ConcurrentHashMap<String, PairID> [*]

Figure 1. Class Diagram - Main Elements

Fig. 1 presents some of the most important classes. The DTSA and Storage classes show the representation of attributes and the other classes in the diagram are a simple sample of elements definition. The Millisecond definition is used to define constraint unit. The elements shown in Fig. 1 are just a sample of some elements defined during modeling. This definition aims to show the level of abstraction represented by the class diagram, which is the representation of attributes and message definition, not focusing on the internal structure of the element, which would be modeled using the state-machine diagram.

The Composite Structure diagram allows to define a detailed view of a classifier structure, the relationship between attributes, input and output interfaces and data flow. In the Etarch architecture, the DTSA is one of the most relevant elements. Therefore, this structure is defined using a Composite Structure diagram. As an entity can be any device, and this behaviour is not relevant to architecture behaviour, then this internal structure will not be modelled.

For the best visualization, the DTSA definition is divided into three pictures, and then the DTSA is divided into two modules, the Resource Adapters that consist of flow control, which are defined as the bases communication protocol. The

Building Block is a composite element responsible of the service control and data storage of protocol data.

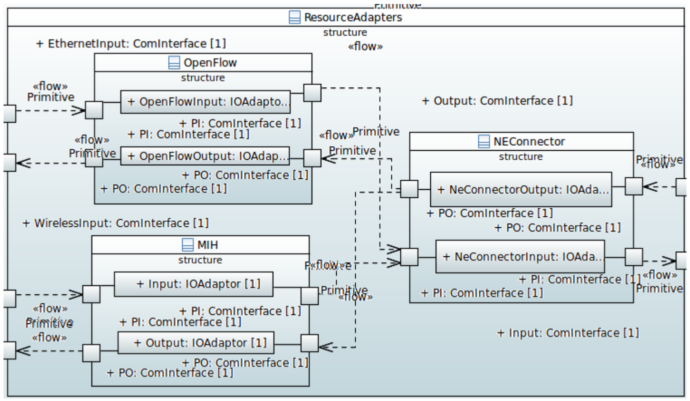


Figure 2. Composite Strucuture Diagram - Resource Adaptors

In Fig. 2, we introduce the structure of the element that represents protocols which standardizes protocol messages of Etarch. In the architecture, this protocol will be used to send messages until a responsible element that will treat and transform the message into architecture ETCP or DTSCP protocol's messages. In this case, the responsible element is called NE Connector.

In Etarch, two protocols are used to standardize the communication: OpenFlow [34] and Media Independent Handover (MIH) [35]. Requests originating from Ethernet will use OpenFlow protocol and requests originating from Wireless will use the MIH standard.

As there is no difference in input and output flow modeling, a distinction has been made in modeling the data flow in the elements. The implementation of elements must represent the concept shown in the modeled structure. As the element shown in Fig. 2 is responsible to intermediate messages in request and response, all input services requests to the DTSA must go through this module.

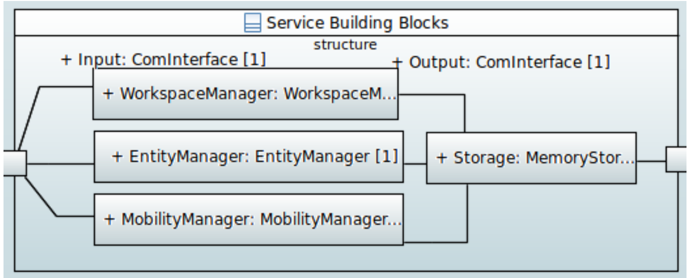

Service Building Blocks
structure
+ Input: ComInterface [1]
+ Output: ComInterface [1]
+ WorkspaceManager: WorkspaceM...
+ EntityManager: EntityManager [1]
+ Storage: MemoryStor...
+ MobilityManager: MobilityManager...

Figure 3. Composite Strucuture Diagram - Building Blocks

The module Building Blocks is depicted in Fig. 3. There are four internal elements in this module. The Workspace Manager that is accountable for all workspace related operations, as creation, attachment, detachment and deletion. The Entity Manager that treats entity requirements, as register and unregister. The Mobility Manager is responsible for mobility operations, as handover among others. The Storage is a generic structure to represent a database, and all the other structures of the same module to finish operation needs to modify the database.

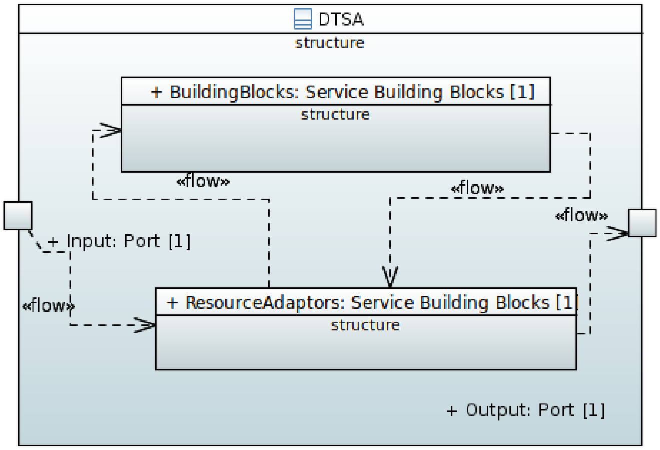

DTSA
structure
+ BuildingBlocks: Service Building Blocks [1]
structure
«flow»
«flow»
«flow»
+ Input: Port [1]
«flow»
+ ResourceAdaptors: Service Building Blocks [1]
structure
+ Output: Port [1]

Figure 4. Composite Strucuture Diagram - DTSA Structure

Fig. 4 represents the relationship in high abstraction level between models presented in Figs. 2 and 3. The request from an Entity must follow what is defined in the DTSA structure. In each module, there are range of behaviours. The most representative behaviours in this paper are Entity Register, Workspace Create, Workspace Attach and Workspace Lookup.

The definition of relevant service behaviour is performed using Sequence Diagrams. The visualization of parameters added in the message hampers visualization when the model is exported to image. Therefore, the name of parameters is added in the name of message.

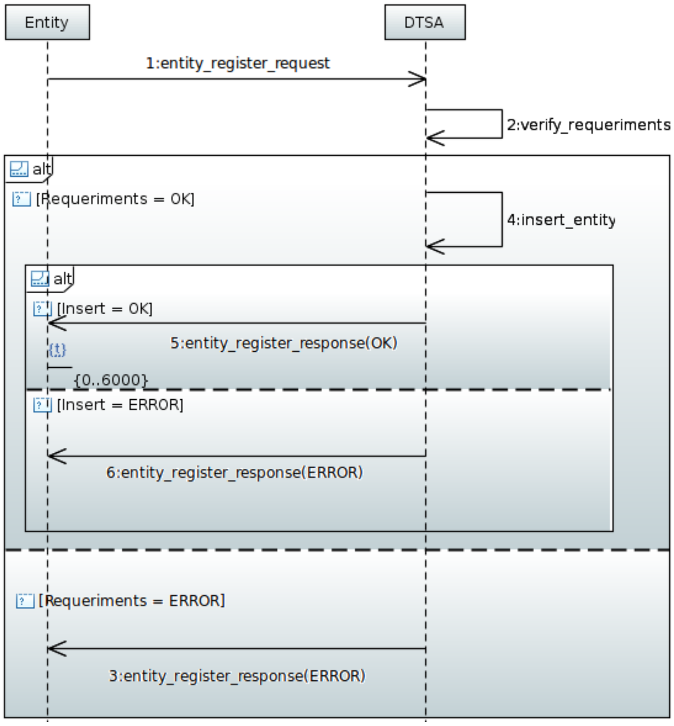


Figure 5. Sequence Diagram -Entity Register

Fig. 5 shows the Sequence Diagram of the entity register service in DTSA. In this diagram, the communication channel is abstracted in such a way that when a call goes to the device DTSA it passes through the flow structure described in the Resource Adaptors. The UML used resources are Time Constraint and Combined Fragment of the "alt" type. The predicates among guards mean the result of called operation.

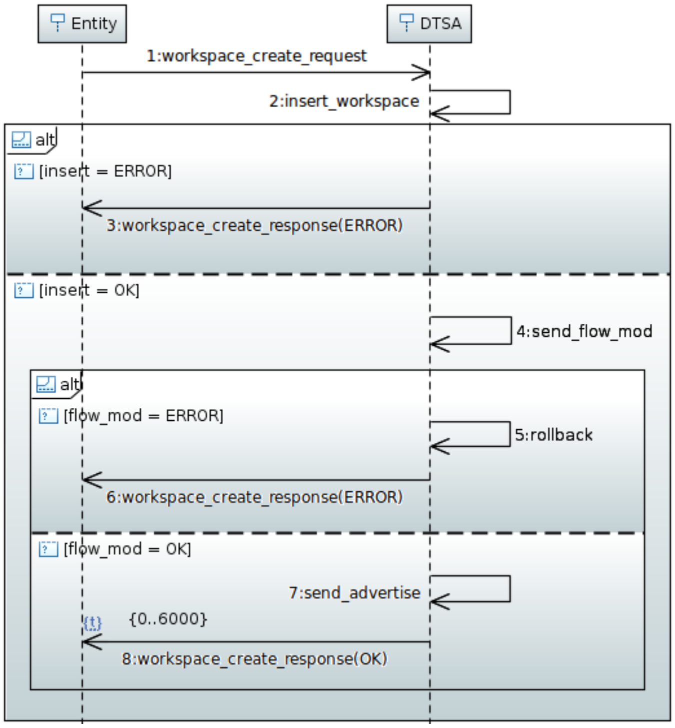


Figure 6. Sequence Diagram - Workspace Create

Fig. 6 represents the flow of workspace creation. In order to execute this operation, Resource Adaptors elements as Building Blocks elements are used. When a workspace is created, its basic information will be saved in storage.

Fig. 7 and Fig. 8 are related. The Workspace Lookup is a sub-process of Workspace Attach. The reference of this diagram is not presented in the figures to improve presentation, but the reference is already in a tool level. In this operation, the time constraints of Workspace Lookup must be taken into account in Workspace Attach. Workspace Lookup has two time constraints, in the search for the next DTSA level the time constraint is essential to the search operation ends.

By proposing an approach of communication protocols modeling, it is important to analyse related works, in particular such one based on state machine. Changes in the modeling language could be tracked by using a mapping function to translate it into a state machine diagram.

This paper presents the first step of a work that aims to create a formalizable scope of UML, UML profiles and enabling enhance the modeling of communication protocols, i.e., it aims to define a set of elements that we can apply transformation rules to a validatable method, or even create such a method. Certainly, it will be necessary to use resources of models transformation between different modeling languages. The Sequence Diagrams, despite of being a little explored approach in this context, are visually more representatives than others modeling techniques. For a formal validation of Sequence Diagrams, we envisage the use of approaches such as transformation into Petri Nets [36].

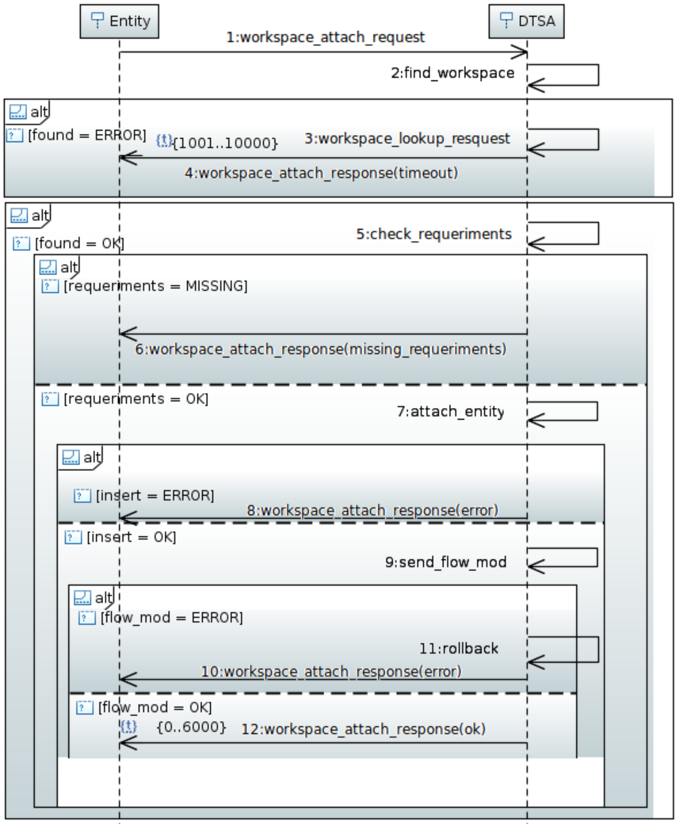


Figure 7. Sequence Diagram - Workspace Attach

Still in the perspective of model transformation, the approach of behavior modeling of communication protocols through Sequence Diagrams services can use synchronization techniques between Sequence Diagrams and other diagrams [37][38].

The modeling of ETArch Protocol, by UML language, follows two ways, being the first one structural modeling, by involving the use of Class and Composite Structure diagrams. The second, through the Sequence diagrams, thus, the necessary elements are: lifeLines, synchronous and asynchronous messages, combined fragments (alt and loop), time and duration observation. Through these elements there are two approaches for transforming models in Petri Nets, in [36] is possible to transform messages, lifelines and combined fragments, however, an approach for modeling observation time and duration is not displayed. Ribeiro and Fern [39] introduced and explained an approach that supports the transformation selected elements in Coloured Petri Nets.

V. Conclusion

This work has shown the modeling of a DTSCP and ETCP communication protocols using the UML language. The UML language has many resources to model components structure, which helps describing the high level of an architectural view. However, the language does not provide a formal definition to communication channel. Therefore, behaviour modeling can

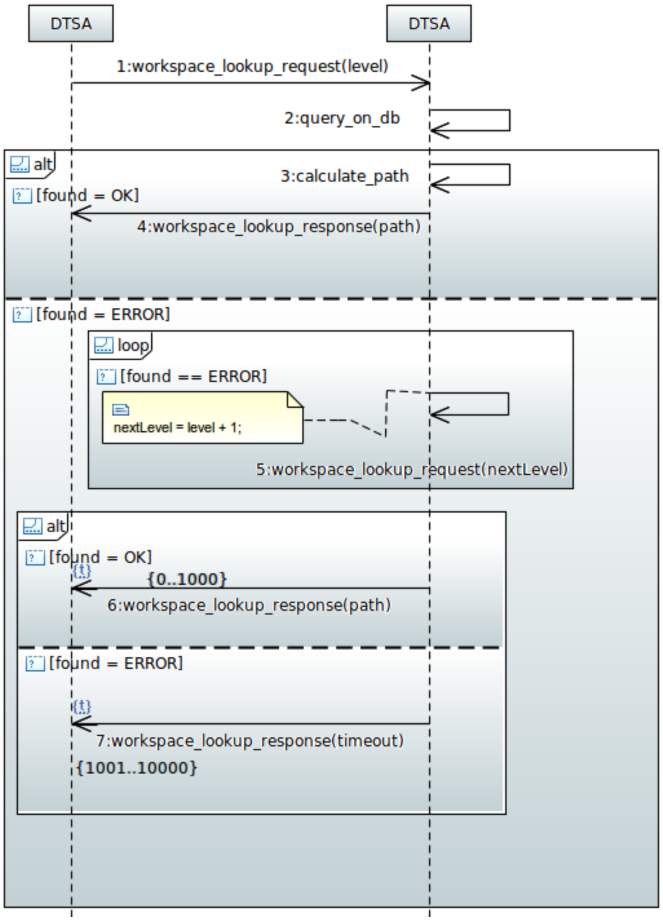

DTSA
DTSA
1:workspace_lookup_request(level)
2:query_on_db
alt
3:calculate_path
[found = OK]
4:workspace_lookup_response(path)
[found = ERROR]
loop
[found == ERROR]
nextLevel = level + 1;
5:workspace_lookup_request(nextLevel)
alt
[found = OK]
{t}
{0..1000}
6:workspace_lookup_response(path)
[found = ERROR]
{t}
7:workspace_lookup_response(timeout)
{1001..10000}

Figure 8. Sequence Diagram - Workspace Lookup

not represent bandwidth constraints, such as limitation can change the behaviour of time constraints.

The time constraints of UML allow the definition of different types for minimum and maximum time, however, it does not allow the creation of relationship between units of measure. The use of combined fragments is limited in static values in predicates.

A limitation of the UML language to define communication protocols is related to definitions of scenarios, because to do this it is necessary to deal with more than one flow definition. In one hand, this possibility is an advantage, but on the other hand its possible definition of data flow is not coincident. For example, in the definition of the Composite Structure Diagram, the data flow is from attribute A to B, and in the Sequence Diagram it is possible to define a message from B to A, injuring the previous definition.

Many ideas can be explored for future work. According to the resources used in this work, it is possible to think about automatic transformation of sequence diagrams to Petri Nets [39] [36] with the purpose of providing formal verification of models. In [36], the great advantage is the use of transformation to a simple Petri Net. However, this transformation is not enough to present the architectural structure defined in the element that is performing the actions. It is necessary a method to transform and attach all related diagrams, and to do this we must define a data flow validation. Therefore, it is possible to model the Etarch protocol services using UML language and the model is able to represent behaviour, time constraints and an abstract architecture of involved elements. The main disadvantage is that it is not possible to validate the complete model, taking into account all structures.

References

[1] “Information Processing Systems - Open Systems Interconnection-’ESTELLE- A Formal Description Technique Based on tan Extended State Transition Model,” 1988.

[2] G. von Bochmann, “Finite State Description of Communication Protocols,” Computer Networks, vol. 2, 1978, pp. 361–372.

[3] S. Simonak, S. Hudak, and S. Korecko, “Protocol Specification and Verification Using Process Algebra and Petri Nets,” in Computational Intelligence, Modelling and Simulation, 2009. CSSim ’09. International Conference on, 2009, pp. 110–114.

[4] O. Ganea, F. Pop, C. Dobre, and V. Cristea, “Specification and Validation of a Real-Time Simple Parallel Kernel for Dependable Distributed Systems,” in Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on, 2012, pp. 320–325.

[5] J. Liu, X. Ye, and J. Li, “CP-Nets Based Methodology for Integrating Functional Verification and Performance Analysis of Network Protocol,” in 11th ACIS International Conference on Software Engineering Artificial Intelligence Networking and Parallel/Distributed Computing (SNPD), June 2010, pp. 41–46.

[6] M. Yusufu and G. Yusufu, “Comparative Study of Formal Specifications through a Case Study,” in International Conference on Information Science and Technology (ICIST), March 2012, pp. 318–321.

[7] S.-S. Park and N. Shiratori, “Distributed Systems Management Based On OSI Environment: Problems, Solutions, and Their Evaluation,” in IEEE 13th Annual International Phoenix Conference on Computers and Communications, 1994, pp. 384–.

[8] J. Day and H. Zimmermann, “The OSI Reference Model,” Proceedings of the IEEE, vol. 71, no. 12, Dec 1983, pp. 1334–1340.

[9] E. D. S. Santos, F. S. F. Pereira, J. H. de Souza Pereira, L. C. Theodoro, P. F. Rosa, and S. T. Kofuji, “Meeting Services and Networks in the Future Internet,” in Future Internet Assembly, ser. Lecture Notes in Computer Science, J. Domingue, A. Galis, A. Gavras, T. Zahariadis, D. Lambert, F. Cleary, P. Daras, S. Krco, H. Muller, M.-S. Li, H. Schaffers, V. Lotz, F. Alvarez, B. Stiller, S. Karnouskos, S. Avessta, and M. Nilsson, Eds., vol. 6656. Springer, 2011, pp. 339–350.

[10] M. Wu and M. Loui, “Modeling Robust Asynchronous Communication Protocols with Finite-State Machines,” IEEE Transactions on Communications, vol. 41, no. 3, 1993, pp. 492–500.

[11] W. Hua, X. Li, Y. Guan, Z. Shi, L. Dong, and J. Zhang, “Formal Verification for SpaceWire Communication Protocol Based on Environment State Machine,” in 8th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM), Sept 2012, pp. 1–4.

[12] R. Alur and D. L. Dill, “A Theory of Timed Automata,” Theoretical Computer Science, vol. 126, no. 2, 1994, pp. 183–235.

[13] X. Wu, H. Ling, and Y. Dong, “On Modeling and Verifying of Application Protocols of TTCAN in Flight-Control System with UPPAAL,” in International Conference on Embedded Software and Systems, 2009, pp. 572–577.

[14] O. Al-Bataineh, T. French, and T. Woodings, “Formal Modeling and Analysis of a Distributed Transaction Protocol in UPPAAL,” in 19th International Symposium on Temporal Representation and Reasoning (TIME), 2012, pp. 65–72.

[15] C. Baier, N. Bertrand, P. Bouyer, and T. Brihaye, “When Are Timed Automata Determinizable?” in Automata, Languages and Programming, ser. Lecture Notes in Computer Science, S. Albers, A. Marchetti-Spaccamela, Y. Matias, S. Nikoletseas, and W. Thomas, Eds. Springer Berlin Heidelberg, 2009, vol. 5556, pp. 43–54.

[16] K. Saleh, “Synthesis of Communications Protocols: An Annotated Bibliography,” SIGCOMM Comput. Commun. Rev., vol. 26, 1996, pp. 40–59.

[17] N. A. Anisimov and M. Koutny, “On Compositionality and Petri nets in Protocol Engineering,” in PSTV, ser. IFIP Conference Proceedings, P. Dembinski and M. Sredniawa, Eds., vol. 38. Chapman & Hall, 1995, pp. 71–86.

[18] C. Lakos, J. Lamp, C. Keen, and B. Marriott, “Modelling Network Protocols with Object Petri Nets,” in Proc. of Workshop on Petri Nets Applied to Protocols. Springer-Verlag, 1995, pp. 31–42.

[19] K. El-Fakih, H. Yamaguchi, G. v. Bochmann, and T. Higashino, “Protocol Re-synthesis Based on Extended Petri Nets,” 2000.

[20] A. Masri, T. Bourdeaud’huy, and A. Toguyeni, “Network Protocol Modeling: A Time Petri Net Modular Approach,” in 16th International Conference on Software, Telecommunications and Computer Networks, 2008, pp. 274–278.

[21] C. Kant, T. Higashino, and G. V. Bochmann, “Deriving Protocol Specifications from Service Specifications Written in LOTOS,” Distrib. Comput., vol. 10, no. 1, 1996, pp. 29–47.

[22] M. Kapus-Kolar, “Comments on Deriving Protocol Specifications from Service Specifications Written in LOTOS,” Distributed Computing, vol. 12, 1999, pp. 175–177.

[23] T. Bolognesi and E. Brinksma, “Introduction to the ISO Specification Language LOTOS,” Comput. Netw. ISDN Syst., vol. 14, no. 1, 1987, pp. 25–59.

[24] G. Booch, J. Rumbaugh, and I. Jacobson, The Unified Modeling Language User Guide, (Addison-Wesley Object Technology Series), A.-W. Professional, Ed. Addison-Wesley Professional, 2005.

[25] C. Lange and M. Chaudron, “An Empirical Assessment of Completeness in UML Designs,” in Proc. Conf. Empirical Assessment in Software Engineering, 2004, pp. 111–121.

[26] M. Jaragh and I. Saleh, “Protocols Modeling using the Unified Modeling Language,” in Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology, vol. 1, 2001, pp. 69–73.

[27] K. Sekaran, “Development of a Link Layer Protocol using UML,” in International Conference on Computer Networks and Mobile Computing, 2001, pp. 309–315.

[28] A. Bagnato, A. Sadovykh, E. Brosse, and T. E. Vos, “The OMG UML Testing Profile in Use–An Industrial Case Study for the Future Internet Testing,” 15th European Conference on Software Maintenance and Reengineering, vol. 15, 2013, pp. 457–460.

[29] F. de Oliveira Silva, M. Goncalves, J. de Souza Pereira, R. Pasquini, P. Rosa, and S. Kofuji, “On the Analysis of Multicast Traffic Over the Entity Title Architecture,” in 18th IEEE International Conference on Networks (ICON), 2012, pp. 30–35.

[30] Open Networking Foundation, “Software-Defined Networking: The New Norm for Networks,” 2012. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf

[31] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, “OpenFlow: Enabling Innovation in Campus Networks,” SIGCOMM Comput. Commun. Rev., vol. 38, no. 2, 2008, pp. 69–74, ACM ID: 1355746.

[32] H. Kim and N. Feamster, “Improving Network Management with Software Defined Networking,” IEEE Communications Magazine, vol. 51, no. 2, 2013, pp. 114–119.

[33] S. T. K. Flavio Oliveira Silva, Joao Henrique. S. Pereira and P. F. Rosa, “Domain Title Service for Future Internet Networks,” in SBRC WPEIF, 2011.

[34] B. Sonkoly, A. Gulyas, F. Nemeth, J. Czentye, K. Kurucz, B. Novak, and G. Vaszkun, “On QoS Support to Ofelia and OpenFlow,” in European Workshop on Software Defined Networking (EWSDN), 2012, pp. 109–113.

[35] D. Griffith, R. Rouil, and N. Golmie, “Performance Metrics for IEEE 802.21 Media Independent Handover (MIH) Signaling,” Wirel. Pers. Commun., vol. 52, no. 3, 2010, pp. 537–567.

[36] M. S. Soares and J. Vrancken, “A Metamodeling Approach to Transform UML 2.0 Sequence Diagrams to Petri Nets,” in Proceedings of the IASTED International Conference on Software Engineering, 2008, pp. 159–164.

[37] J. Whittle and J. Schumann, “Generating statechart designs from scenarios,” in Software Engineering, 2000. Proceedings of the 2000 International Conference on, 2000, pp. 314–323.

[38] R. Grønmo and B. Møller-Pedersen, “From sequence diagrams to state machines by graph transformation,” in Proceedings of the Third International Conference on Theory and Practice of Model Transformations, ser. ICMT’10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 93–107.

[39] O. R. Ribeiro and J. M. Fern, “Some Rules to Transform Sequence Diagrams into Coloured Petri Nets,” in In 7th Workshop and Tutorial on Practical Use of Coloured Petri Nets and the CPN Tools, 2006, pp. 237–256.

Multicast Traffic Aggregation through Entity Title Model

Maurício Amaral Gonçalves, Flávio de Oliveira Silva,
João Henrique de Souza Pereira and Pedro Frosi Rosa
Federal University of Uberlandia
Uberlandia - Minas Gerais - Brazil
Email: mauricioamaralg@gmail.com, flavio@facom.ufu.br,
joaohs@ufu.br, frosi@facom.ufu.br

***Abstract*—Internet was designed in a totally different context than the one existing today. New applications have brought a new set of requirements which were not properly resolved due to architectural limitations. Therefore, the Internet architecture must be reviewed in a *clean slate* approach. In this context, Entity Title Model represents a revolutionary way to semantically understand the entities, observing their needs and capabilities in order to better serve them, through a new flexible architecture with several innovations, especially in addressing and routing aspects. This paper presents a protocol capable of providing efficient multicast at the network layer, based on ETArch over OpenFlow. Multicast is an important requirement for applications involving the transmission of multimedia content, real-time communication and data-sharing services. We describe some experiments and present a comparison between a video application, first implemented using TCP/IP with unicast and multicast services, and then using ETArch focusing on multicast traffic aggregation. The results showed that the bandwidth consumption using our architecture remains constant just as the traditional one; however, our approach uses slightly less bandwidth, provides better strategies for the control plane, improves the group addressability, and facilitates its deployment based on the broad support to Openflow by leading equipment suppliers.**

***Keywords*—telecommunications networks; Internet; multicast; future Internet; clean slate; entity title model.**

I. Introduction

The main concepts of Internet were designed in the sixties [1], and its core protocols were created in the early seventies [2]. If, on one hand, the stability of these protocols led to the the popularity of the Internet, on the other hand, they now refrains its modernization [3]. After four decades and a huge success, much of the initial design of the Internet is still in place. However, applications vastly different from those that initially used the network are now being deployed, bringing a new set of requirements, such as multicast, which current Internet is not able to satisfy in a proper way due to its limitations [4].

Multicast is the ability to deliver data to a group of target entities simultaneously in a single transmission. This aspect is closely linked to how addressing occurs and what routing algorithms are used to reach the entities over the network. The main problem of *Internet Protocol* (IP) addressing is in its ambiguous addressing, which represents both location and identification [5]. This limitation prevents the addressing of a multicast group natively, because there is no unique physical location for a multicast group, and so, the IP address could not be used to locate the members. IP Multicast [6] skirted this problem by using specific reserved address blocks and an implementation of data replication in routers, which became responsible for maintaining the multicast groups. Given the complexity and limitations of this approach, the IP multicast is still not widely used today, even after twenty years of its conception [7].

Researchers from all over the world are engaged in the design of a new Internet from scratch. The *clean slate* approach frees the research from the legacy and fosters innovations [8]. One approach that has taken power in recent years is the *Software-Defined Networking* (SDN), designed in a partnership between UC Berkeley and Stanford University. The Software Defined Networks represents a milestone for advanced researches on new architectures of computer networks. The decoupling between control plane and data plane in network devices contributed with the arising of numerous research projects that collaborated to get the SDN level of maturity as it is today.

Entity Title Architecture (ETArch) presents a vision of how entities are enabled to semantically specify their requirements and capabilities, in order to establish a communication between two or more entities, using a naming scheme based *titles*, which are topology independent and unambiguously designations, and new approaches for addressing and routing aspects [9]. In this work, the ETArch implementation was based on Openflow [10], and focuses on multicast capability, but is not limited to this approach or to this requirement.

The remainder of this paper is organized as follows: Section II describes the related work; Section III presents the Entity Title Architecture; Section IV details the implementation; Section V describes the experiment; Section VI discusses the results obtained and Section VII presents some concluding remarks and potential future works.

II. Related Work

SDN [11][12] represents an extraordinary opportunity to rethink computer networks. It consists of an abstraction that separates the software that controls the network elements from the forwarding plane, providing an open and well-defined interface to control and modify the behavior of network at runtime.

ISBN: 978-1-61208-360-5

The Future Internet subject is benefited by the range of possibilities offered by the SDN in various applications. In [13], for example, the *Border Gateway Protocol* (BGP) gets an important reinforcement from a *Routing Control Platform* (RCP) system-based, also controlled by SDN. Alternatives to support different applications requirements, such as delivery guarantee, appears in contrast to traditional TCP/IP, as presented by *Dias et al.* [14].

In the state of the art of the SDN, there is an increase in the number of network elements that support OpenFlow; however, although SDN has brought to light the possibility of inferring in the network programming behavior, this is not an easy task. The researchers are engaged in creating software able to abstracting the various features controlled by the network, such as *Foster et al.* [15] and *Kim and Feamster* [16], which offer important contributions for the advancement of researches in this area.

In the EU (European Union), about a hundred different projects are funded under the Seventh Framework Programme (FP7), and some of which are directly related to the Future Internet as 4WARD, CHANGE, MEDIEVAL, PURSUIT, SAIL, SENSEI, TRILOGY and UNIVERSELF [17]. These projects work with different aspects of future networks, and many of them present *clean slate* approaches.

The 4WARD *Netinf* [18] presents an information-centric networking paradigm, based on a distributed system over the network, which controls the communication and provides useful services, such as caching, storage and transporting. It uses a naming scheme independent of the network, called *Identifier*, which is related to *Title* presented at this work. These identifiers are used to register and resolve *Information Objects*, which are primitives exchanged during communication.

In the United States, the *Future Internet Architectures* (FIA) [20], which represents a consolidation of the previous program contains four projects that currently are dealing with aspects of the network, such as content-centric networks, mobility, cloud computing and security. The MobilityFirst [21] network architecture focuses on mobility and propose new protocol stack that considers a new naming scheme based on *Globally Unique Identifier* (GUID) that can provide mobility and multicast. The *Title* is related with the GUID, but the concept of workspace provides a out-of-band control for packet delivery, while in MobilitFirst the control happens in-band.

The IP Multicast, proposed by *Deering* [6], presents limitations both in technical and business aspects [22], such as: limited number of multicast addresses, inability of managing groups dynamically, security constraints, complex architecture, and difficulties in deployment and management.

In IPv6, the concept of broadcast addresses was replaced by the multicast addresses [23]. Furthermore, the network interfaces became able to join different multicast groups. This architecture provides dynamic IP address allocation [24], which can be defined in different scopes [25].

The multicast based on IPv6 presents challenges regarding security [26], with vulnerabilities that can be exploited by attacks. Moreover, scenarios with mobility requirements, where users share frequencies with limited bandwidth, present a number of challenges [27], fueled by the combination of these two requirements.

Due to these limitations, the deployment of IP Multicast occurs slowly [28], which promoted the adoption of the *Application Layer Multicast* (ALM) [29], also known as *End System Multicast* (ESM), in which most of the issues of multicast over IP are addressed at the application layer, facilitating its adoption by not implying changes in the network architecture.

The ability to easily deploying the ALM protocol is a great advantage compared to IP Multicast, which in other hand provides a better optimization of communication bandwidth, partially wasted in ALM due to its multicast strategy, which is based on packet replication over the distribution trees [29]. Moreover, even using ALM, issues such as mobility presents several challenges due to limitations imposed by the architecture.

In this scenario, with different designs, the Entity Title Architecture is an additional proposal that may contribute to this area of research. The outlook presented supports the main ideas about this work, which are: a new protocol stack for the Internet replacing TCP/IP stack, a new naming and addressing scheme, an experimental approach using SDN, an implementation of real multicast, and a vision for collaboration between research community.

III. Entity Title Architecture

ETArch is a *clean slate* approach for the Future Internet, which proposes: a separation of responsibilities between the data and control planes, a semantic proximity of the layers, and a new strategy to addressing and routing. It works as an intermediary layer, as shown in Figure 1. To properly understand how this architecture works, it is first necessary to understand a few concepts:

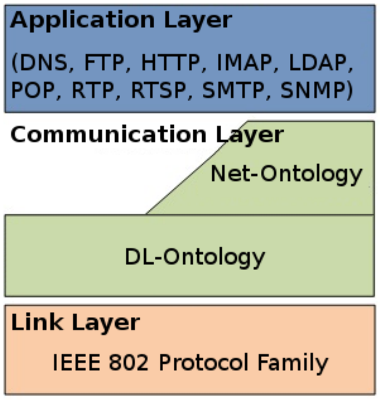


Figure 1. ETArch Stack.

- Entity: is a thing with communication requirements which can be semantically understood from top to bottom layers. Some examples: a content, a service, a sensor device, pad or smart phone, a user, an application, a system, a process. The entity has some titles, requirements and a variable location over time.

- Title: is a designation to ensure an unambiguous identification of an entity. One title designates only one entity, but one entity may have more than one title. The title plays a key role in order to provide the horizontal addressing entities.
- Requirements: are needs defined in the establishment of the logical link (workspace), which represent also the capabilities that entities must support to make part of a communication.
- Capabilities: are features supported by entities in order to meet the communication needs for a particular purpose.
- Horizontal Addressing: is an addressing scheme independent of the physical location of network entities, without the need for bandwidth reservation, network segmentation or specific physical connections.
- *Domain Title Service* (DTS): consists of a distributed system over the network [30], responsible for the maintenance of entity and provisioning of logical links required for communication. It is also able to understand their capabilities and needs, and for providing of features to treat them properly. Comprising *Domain Title Service Agents* (DTSA), it plays an important role in key aspects of the network, such as names and addresses, and have the ability to share the connection between the communicating entities. Throughout the network, DTSA are distributed in that domain being deployed at servers and network elements (switches, routers, and so on).
- Workspace: is a logical bus that has a title and contains network elements required to support the communication of the entities. The workspace is created by an entity that wants to communicate with a specific purpose. During its inception, the entity informs the set of requirements that must be supported by all entities who want to be part of the workspace. A new entity can be associated with an existing workspace and, if so, the logical bus can be extended to handle your communication. Likewise, an entity can move through the DTS being able to maintain it communicating. The main concept introduced by workspace is that the destination address is its title. Another important concept is that primitive, for example a stream, is sent once by the source and can be received by all the entities sharing it.
- DL-Ontology: is a logical link layer, able to semantically interpret and meet the requirements of the upper layers, using the infrastructure of the network optimally. It is the realization of logical link concept, being responsible for delivering data to the entities that compose the workspace.
- Net-Ontology: is responsible to semantically interpret the needs of the entities, and implement them through the DL-Ontology layer. It is a mechanism for semantic reasoning and features modularization, which links requirements and capabilities, establishing communication according to entities needs.
- *Entity Title Control Protocol* (ETCP): is a protocol that defines the communication between entities and DTS. It provides maintenance services of the entity and management of workspaces services. Example: entity-register, workspace-attach.
- *Domain Title Service Control Protocol* (DTSCP): is a protocol that defines the communication between DTSA's. Provides workspace search and register inter-DTSA services. Example: workspace-register and workspace-lookup.

One the main points of ETArch is the *horizontal addressing*, which solves the problem of ambiguity between identification and localization of the current architecture. In this approach, the identification of the *entity* is defined by its *titles*, and its localization is controlled by the *DTS* Agent immediately superior. When an *entity* wants to communicate, it creates a *workspace* by sending an ETCP message to DTSA. This *workspace* has a set of *requirements* which must match with the *capabilities* of the *entities* that wants to communicate. All the data transmitted over the network is delivered by the DL-Ontology, which is the main protocol of this architecture. It may be necessary to perform some additional processing by the network elements and hosts during the interpretation of the Net-Ontology, which defines the communication requirements. All communication is orchestrated by DTS, which is a distributed system materialized by their agents that communicate via DTSCP protocol.

IV. Implementation

This section aims to present an implementation scenario as a *proof of concepts*, regarding *workspace* concept applied to achieve the goal of *multicast aggregation*. We are mainly interested in observing the behavior of the network in the face of features like multicast, provided naturally by the architecture.

In order to overcome the existing limitations in the TCP/UDP/IP, including underlying protocols such as Ethernet and others, we developed a network interface which provides for the entities in a distributed environment free from legacy Internet protocols.

ETArch proposes a division between data and control planes, as well as OpenFlow, and its main components are: the DL-Ontology and Net-Ontology layers (in the data plane), and the DTS with its agents (in the control plane). The following sections describe how these components were implemented.

A. Net/DL-Ontology

The implementation has four main modules designed to have high cohesion and loose coupling for the entire workspace enrollment project.

The Ontology module is responsible for the design Title Model including the concepts of: DTS, Workspace, DL-Ontology, Title and Entities. It was modeled with software Protégé [31], and generated in *Ontology Web Language* (OWL) by using *OWL API* [32].

The module responsible for interpretation of OWL is under construction by the use of Jena. The reasoning of the ontology is a central point of the semantic approach, since it makes possible the creation of inference rules to implement the intelligence of Title Model. At this step, a parser based on regular expressions was used, and in the next stage of implementation will be included the reasoning.

The interface module is the implementation of a Java API for use by the parser and reasoning module. This includes communication through Raw Socket API, built in C language.

The Physical Medium Access module is responsible for the communication with physical layer allowing the primitive DL-Ontology to be sent to the physical environment without Internet protocols, such as Ethernet, IP, TCP, UDP, or SCTP.

B. DTSA (as an Openflow Controller)

As the DTSA's task of coordinating network elements is closely related to that of managing flows by an OpenFlow controller, we have decided to implement the first on top of the latter. In a nutshell, we extended the FloodLight open-source OpenFlow controller [33] to closely work with the DTSA.

The extensions to the Floodlight controller consisted in a new module that instantiates the DTSA and handles the exchange of DTS control messages.

As a extented *IOFMessageListener*, this module is able to listen incoming messages. By default, all messages that do not match any of the rules in the switch flow table are sent to the DTSA. When a message is received, the listener is called and checks if the message is a defined primitive. If so, the message is delivered to the DTSA which process it and modify the switches using a *flow_mod* [34].

V. Experiments

To experiment and evaluate the Entity Title Architecture, especially the workspace concept with its multicast capabilities, we conducted some experiments.

A simple topology, as shown in Figure 2, was defined. On the right side, a server contains a video application that produces a flow-based *Motion Joint Photographic Experts Group* (MJPEG). On the left side, at a host, one or more clients where instantiated during the experiments. Between the the hosts are three OpenFlow switches. Entities hosted at any host, including DTS Agent, are able to send and receive DL-Ontology primitives. Although, it is a simple topology that reflects a common situation where a server and a client are separated by a set of switches. The topology was created using Mininet [35], a system for rapid prototyping of OpenFlow-based networks.

To compare the Entity Title Architecture and the use of TCP/IP architecture for the networking, three different server applications where created. The first and the second ones based on UDP and IP protocols with *unicast* and *multicast* approaches respectively, and the third one based on our approach. Essentially, these applications are the same, and the main difference between them is just the way sockets are created and used.

At the application layer, a *Real-time Transport Protocol* (RTP) [36] based message is created, then, in the first case, Datagram Socket is used to send this message. The second video application, that uses the workspace, creates a *Finsocket*, which is based on Raw Sockets. Raw Sockets does not use the TCP/IP stack and directly creates a frame and send it over the physical medium. In fact, the *Finsocket* does create a frame based on the Ethernet frame, but it does not contain the traditional information in its headers. Instead, the source address contains the leftmost bits of the workspace title and the destination address field is the rightmost bit.

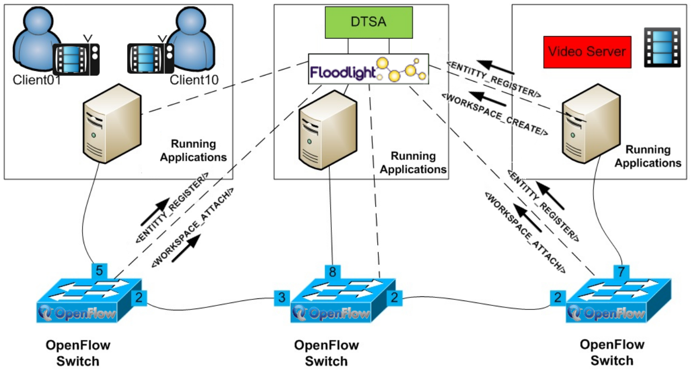
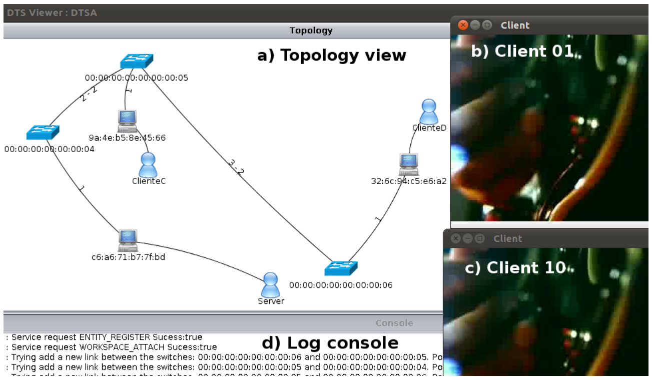


Figure 2. Scenario used for experimentation.

Additionally, a management application for the DTSA was conceived, to allow a better visualization of the proposed scenario, as shown in Figure 3. Also, in this figure, one can observe two video subscribers attached to the workspace, and so, receiving the same flow.



```
: Service request ENTITY_REGISTER Sucess:true
: Service request WORKSPACE_ATTACH Sucess:true
: Trying add a new link between the switches: 00:00:00:00:00:00:00:06 and 00:00:00:00:00:00:00:05. Po
: Trying add a new link between the switches: 00:00:00:00:00:00:00:05 and 00:00:00:00:00:00:00:04. Po
```

Figure 3. DTSA Management Application and Clients attached to the workspace.

VI. Results

At the experiments, a server application has been started and a different number of clients connected to it, requesting data. Considering the UDP/IP Unicast server application, in proportion as the number of clients grows, there is also an increase in bandwidth usage caused by the data replication on various flows instantiated. The video server that uses the Entity Title Architecture remains with a constant use of bandwidth at the source, no matter the number of clients. This is because the data is sent to workspace and a client connects to it, not directly to the server. The same occurs with the application using the UDP/IP Multicast, because the IP Multicast groups concept is related to ETArch workspace. In both approaches the data is replicated in the network elements; however, the IP Multicast has problems that make it unfeasible in global proportions for practical purposes [37]. Figure 5 shows the use of the bandwidth obtained from the comparison between the IP Unicast, IP Multicast and DL-Ontology approaches.

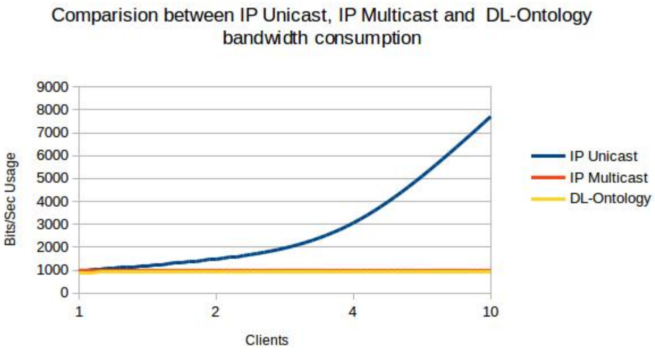


Figure 4. Bandwidth usage at the source versus the number of clients between all tested approaches.

Figure 5 focus on IP Multicast and DL-Ontology comparison. The results are similar; however, the DL-Ontology approach uses slightly less bandwidth, given the change of protocols used in the network and transport layer.

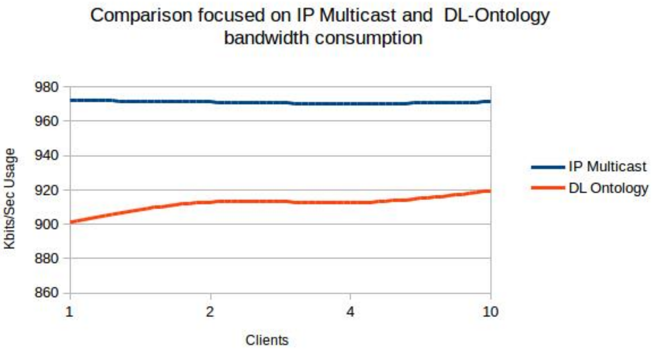


Figure 5. Bandwidth usage at the source versus the number of clients between IP Multicast and DL-Ontology approaches.

Although the results between the proposed and the conventional *multicast* are similar, the major advantage of our proposal is the possibility of deployment on a global scale by taking advantage of SDN. The ETArch approach proposes a new model for the Internet that gives natural support for *multicast* communication through drastic changes in aspects of addressing, identification and routing. The IP Multicast has limitations at: addressing, because of the limited number of *multicast* addresses, restricted to *class D*; network supporting, since it is necessary that all core devices provide this service; and control signaling, which imposes an impractical overhead in global scales.

VII. Conclusion

Considering the new set of requirements, the Internet architecture must be revised. This review process using a *clean slate* can free researchers from current deficiencies by providing a rich environment for experimentation.

In this article, we presented a SDN-based implementation of the *Entity Title Architecture*, and its application to address the multicast requirement. This work focused on the presentation of the main concepts of the architecture, demonstrating that the aggregation of multicast becomes a trivial task, because it is something intrinsic to the architecture.

Although OpenFlow can be used to implement the new naming, routing and addressing schemes, the literature on the topic does not contain detailed descriptions of how this can be done and this work aims to contribute in this matter too. So in addition to experimentally demonstrate the Tile Entity Architecture, this works also shows how an IP centered OpenFlow switch, compatible with OpenFlow 1.0 specification, can be used in networks that completely drop the TCP/IP stack from the data plane using a new semantics the for flow table.

The evaluation of the implemented architecture showed that the bandwidth used for the source remains constant regardless of the number of customers connected to it. The impact of this fact is that real connections can be used to provide services, such as high definition videos with efficient power consumption.

This was an expected result, because the Entity Title Architecture is based on a new naming and addressing scheme, where the destination address is the workspace and while the packet is sent to it, all entities that are part of it receives this packet bringing the architecture a seamless multicast capability. The workspace also provides mobility, cause it can move between the switchs, and in the presence of this event, the flow table will be automatically updated.

The approach presented in this paper is a more efficient form of communication if compared to the current solutions, such as IP Multicast (at network layer) and ALM (at application layer), by not having the limitations of TCP/IP architecture as demonstrated in this work. The ETArch provides a real *multicast* by drastic changes in routing and addressing schemes. There is no data repetition in the communication within the workspace, cause it provides a natural support to that requirement, differently from the ALM, which despite reduces the replication level, does not eliminate it completely, by presenting a strategy that does not take into account the access and distribution elements, just the core elements. Unlike IP Multicast, in ETArch approach it is possible for a host be attached to more than one workspace at the same time, through the flexibility in the routing rules provided by this architecture.

We are currently working on improving the security, routing and control plane aspects, which should be subject of the future work.

The results show that we are facing a viable approach to bring richer and more efficient services to the network, collaborating with research aimed to define, design and implement the next generation of computer network architectures.

Acknowledgment

This project was made under the support and belief of MEHAR team, valuable working partners. Therefore, we would like to thank all who contributed in any way to complete this stage of our research.

References

[1] P. Baran, "On distributed communications networks," *IEEE Transactions on Communications Systems*, vol. 12, no. 1, pp. 1–9, Mar. 1964.

[2] V. Cerf and R. Kahn, "A protocol for packet network intercommunication," *Communications, IEEE Transactions on*, vol. 22, no. 5, pp. 637–648, 1974.

[3] J. H. De Souza Pereira, S. T. Kofuji, and P. F. Rosa, "Distributed systems ontology," in *New Technologies, Mobility and Security (NTMS), 2009 3rd International Conference on*, 2010, pp. 1–5.

[4] T. Zahariadis *et al.*, "Towards a future internet architecture," in *The Future Internet. Future Internet Assembly 2011: Achievements and Technological Promises*, ser. LNCS, J. Domingue, A. Galis, A. Gavras, T. Zahariadis, and D. Lambert, Eds. Berlin, Heidelberg: Springer-Verlag, 2011, vol. 6656, pp. 7–18. [Online]. Available: http://www.springerlink.com/content/978-3-642-20897-3#section=881237&page=15&locus=86 [retrieved: May. 2014]

[5] D. Farinacci, D. Lewis, D. Meyer, and V. Fuller, "The Locator/ID separation protocol (LISP)." [Online]. Available: http://tools.ietf.org/html/rfc6830 [retrieved: May. 2014]

[6] S. Deering, *Host extensions for IP multicasting*, ser. Request for Comments. IETF, Aug. 1989, no. 1112, published: RFC 1112 (Standard) Updated by RFC 2236. [Online]. Available: http://www.ietf.org/rfc/rfc1112.txt [retrieved: May. 2014]

[7] Y.-h. Chu, S. Rao, S. Seshan, and H. Zhang, "A case for end system multicast," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 8, pp. 1456 – 1471, Oct. 2002.

[8] J. Roberts, "The clean-slate approach to future internet design: a survey of research initiatives," *annals of telecommunications - annales des tlcommunications*, vol. 64, no. 5-6, pp. 271–276, May 2009. [Online]. Available: http://www.springerlink.com/content/e240776641607136/ [retrieved: May. 2014]

[9] F. de Oliveira Silva *et al.*, "Semantically enriched services to understand the need of entities," in *The Future Internet*, ser. Lecture Notes in Computer Science, F. lvarez *et al.*, Eds. Springer Berlin / Heidelberg, 2012, vol. 7281, pp. 142–153. [Online]. Available: http://www.springerlink.com/content/1222874ul734676k/abstract/ [retrieved: May. 2014]

[10] N. McKeown *et al.*, "OpenFlow: enabling innovation in campus networks," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, pp. 69–74, Mar. 2008, ACM ID: 1355746.

[11] G. Goth, "Software-Defined networking could shake up more than packets," *IEEE Internet Computing*, vol. 15, no. 4, pp. 6–9, Aug. 2011.

[12] K. Greene, "TR10: Software-Defined networking," *MIT Technology Review*, vol. 112, no. 2, Apr. 2009. [Online]. Available: http://www.technologyreview.com/web/22120/ [retrieved: May. 2014]

[13] C. Rothenberg *et al.*, "Revisiting IP Routing Control Platforms with OpenFlow-based Software-Defined Networks," *XXX SBRC - III Workshop de Pesquisa Experimental da Internet do Futuro-WPEIF*, p. 6, 2012.

[14] A. Dias *et al.*, "Cross Layers Semantic Experimentation for Future Internet," *XXX SBRC - III Workshop de Pesquisa Experimental da Internet do Futuro-WPEIF*, p. 16, 2012.

[15] N. Foster *et al.*, "Languages for Software-Defined Networks," *IEEE Communications Magazine*, p. 128, 2013.

[16] H. Kim and N. Feamster, "Improving Network Management with Software Defined Networking," *IEEE Communications Magazine*, p. 114, 2013.

[17] E. Commission, "The network of the future - projects," http://cordis.europa.eu/fp7/ict/future-networks/projects_en.html, 2012. [Online]. Available: http://cordis.europa.eu/fp7/ict/future-networks/projects_en.html [retrieved: May. 2014]

[18] M. DAmbrosio, M. Marchisio, V. Vercellone, B. Ahlgren, and C. Dannewitz, "4WARD. second NetInf architecture description," 2010. [Online]. Available: http://www.4ward-project.eu/index.php?s=file_download&id=70 [retrieved: May. 2014]

[19] J. H. d. S. Pereira, F. d. O. Silva, E. Lopes Filho, S. T. Kofuji, and P. F. Rosa, "Title model ontology for future internet networks," in *Future Internet Assembly 2011: Achievements and Technological Promises*. Springer-Verlag, 2011, vol. 6656, p. 465.

[20] N. S. Foundation, "NSF future internet architecture project," http://www.nets-fia.net/, 2011. [Online]. Available: http://www.nets-fia.net/ [retrieved: May. 2014]

[21] I. Seskar, K. Nagaraja, S. Nelson, and D. Raychaudhuri, "MobilityFirst future internet architecture project," in *Proceedings of the 7th Asian Internet Engineering Conference*, ser. AINTEC '11. New York, NY, USA: ACM, 2011, pp. 1–3. [Online]. Available: http://doi.acm.org/10.1145/2089016.2089017 [retrieved: May. 2014]

[22] C. Diot, B. Levine, B. Lyles, H. Kassem, and D. Balensiefen, "Deployment issues for the IP multicast service and architecture," *IEEE Network*, vol. 14, no. 1, pp. 78 –88, Feb. 2000.

[23] R. Hinden and S. Deering, *IPv6 Multicast Address Assignments*, ser. Request for Comments. IETF, Jul. 1998, no. 2375, published: RFC 2375 (Informational). [Online]. Available: http://www.ietf.org/rfc/rfc2375.txt [retrieved: May. 2014]

[24] D. Thaler, M. Handley, and D. Estrin, *The Internet Multicast Address Allocation Architecture*, ser. Request for Comments. IETF, Sep. 2000, no. 2908, published: RFC 2908 (Historic) Obsoleted by RFC 6308. [Online]. Available: http://www.ietf.org/rfc/rfc2908.txt [retrieved: May. 2014]

[25] R. Hinden and S. Deering, *IP Version 6 Addressing Architecture*, ser. Request for Comments. IETF, Feb. 2006, no. 4291, published: RFC 4291 (Draft Standard) Updated by RFCs 5952, 6052. [Online]. Available: http://www.ietf.org/rfc/rfc4291.txt [retrieved: May. 2014]

[26] E. Davies, S. Krishnan, and P. Savola, *IPv6 Transition/Co-existence Security Considerations*, ser. Request for Comments. IETF, Sep. 2007, no. 4942, published: RFC 4942 (Informational). [Online]. Available: http://www.ietf.org/rfc/rfc4942.txt [retrieved: May. 2014]

[27] I. Romdhani, M. Kellil, H.-Y. Lach, A. Bouabdallah, and H. Bettahar, "IP mobile multicast: Challenges and solutions," *IEEE Communications Surveys Tutorials*, vol. 6, no. 1, pp. 18–41, 2004.

[28] W.-P. K. Yiu and S.-H. G. Chan, "Offering data confidentiality for multimedia overlay multicast: Design and analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 5, no. 2, pp. 13:1–13:23, Nov. 2008. [Online]. Available: http://doi.acm.org/10.1145/1413862.1413866 [retrieved: May. 2014]

[29] M. Hosseini, D. T. Ahmed, S. Shirmohammadi, and N. D. Georganas, "A survey of application-layer multicast protocols," *Commun. Surveys Tuts.*, vol. 9, no. 3, pp. 58–74, Jul. 2007. [Online]. Available: http://dx.doi.org/10.1109/COMST.2007.4317616 [retrieved: May. 2014]

[30] J. H. de Souza Pereira, S. T. Kofuji, and P. F. Rosa, "Horizontal addressing by title in a next generation internet," in *2010 Sixth International Conference on Networking and Services (ICNS)*. IEEE, Mar. 2010, pp. 7–11.

[31] M. Horridge, M. Musen, C. Nyulas, S. Tu, and T. Tudorache. (2012, May) protg. [Online]. Available: http://protege.stanford.edu/ [retrieved: May. 2014]

[32] M. Horridge, S. Bechhofer, and O. Noppens, "Igniting the OWL 1.1 touch paper: The OWL API." in *OWLED*, vol. 258. Citeseer, pp. 6–7. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.4920&rep=rep1&type=pdf [retrieved: May. 2014]

[33] Big Switch, "Floodlight OpenFlow controller," http://floodlight.openflowhub.org/, Jan. 2012. [Online]. Available: http://floodlight.openflowhub.org/ [retrieved: May. 2014]

[34] F. de Oliveira Silva, M. Goncalves, J. de Souza Pereira, R. Pasquini, P. Rosa, and S. Kofuji, "On the analysis of multicast traffic over the entity title architecture," in *2012 18th IEEE International Conference on Networks (ICON)*, p. 3035.

[35] B. Lantz, B. Heller, and N. McKeown, "A network in a laptop: rapid prototyping for software-defined networks," *Proceedings of the Ninth ACM SIGCOMM Workshop on Hot Topics in Networks*, pp. 19:1–19:6, 2010, ACM ID: 1868466.

[36] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, *RTP: A Transport Protocol for Real-Time Applications*, ser. Request for Comments. IETF, Jul. 2003, no. 3550, published: RFC 3550 (Standard) Updated by RFCs 5506, 5761, 6051. [Online]. Available: http://www.ietf.org/rfc/rfc3550.txt [retrieved: May. 2014]

[37] A. Boudani and B. Cousin, "SEM: a new small group multicast routing protocol," in *10th International Conference on Telecommunications, 2003. ICT 2003*, vol. 1, pp. 450–455.

Design and Development of an Android Accounting Application Using Web Services and Quality of Experience for Mobile Computing

Ustijana Rechkoska
"St. Paul the Apostle" University
Faculty for Computer Science and Engineering
Ohrid, Republic of Macedonia
ustijana@gmail.com

Danco Davcev
"Sts Cyril and Methodius" University
Faculty for Computer Science and Engineering
Skopje, Republic of Macedonia
dancodavcev@gmail.com

Darjan Djamtovski
"St. Paul the Apostle" University
Faculty for Communication Network Security
Ohrid, Republic of Macedonia
dzamtomk@gmail.com

Carlo Ciulla
"St. Paul the Apostle" University
Faculty for ISVMA
Ohrid, Republic of Macedonia
cxc2728@njit.edu

Jordan Sikoski
State University of Tetovo
Tetovo, Republic of Macedonia
jordans@t-home.mk

Abstract—In this work, we propose an easier and innovative way of calculating the salary of institutions' employees, through their smartphones running the *Android OS*. The application is developed in *Android Ice Cream Sandwich* - the most popular version of *Android*. The specifications and features of the smartphones had achieved the highest level of quality, allowing performance of tasks concerning mobile computing. The application proposed in this work allows calculating the salaries of the employees of the institution - Universities staff, proposing an original, efficient way of documentation managing at institutions. Our approach improves the *Quality of Experience (QoE)*, providing the users all necessary resources and performances, which refer to a user-friendly mobile application, using *Web Services*, *QoE* evaluated referring the quality of the graphical user interface, efficiency and time saving, enabling *Cloud connected experience* possibility.

Keywords-Android application development; mobile computing; salary calculation; Web Services; Quality of Experience (QoE)

I. INTRODUCTION

Android, as a *Linux*-based operating system designed for touchscreen devices like smartphones and tablet computers, is an open source and its code is released under the Apache License which allows the software to be modified and distributed by device manufacturers, wireless carriers, as well as individual users [8][9].

The history of *Android* raises through some versions, from 1.0 to 4.4.2, so far. In this work, we have chosen *Android 4.0* because it provides new graphical interface; there is a security concept improvement and it involves the cloud environment connection. One of the main features of *Android 4.0* is the new *connected Cloud experience*, *Android Beam, refined User Interface (UI)*, security for applications, content and enhancement for enterprise *Virtual Private Network (VPN)* client *Application Programming Interface (API)*. The design of this application is developed consistently with the new *refined UI* which allows easier usability and easier transition of the salary calculation. This application is connected to the *Web Server,* which allows dynamic download of the necessary information for calculation of the salary. The features of this application are described in the third section of this work.

In the next section, the performed tests and results are presented, continuing at the fifth section of this work with *Quality of Experience (QoE)* evaluation concerning mobile computing.

Our approach for managing the documentation of an institution provides saving time, energy and money. It provides comparison to previously developed *Android* applications and gives the reasons why this application is more suitable for usage. This work provides the features of the application and how they can be used. The results of the tests done to the application are presented. The conclusion and our concept for working on this type of application are given. The *Cloud connected experience* allows users to synchronize: photos, e-mails, applications, and contacts [2]. *Android 4.0* provides easier implementation of the applications to manage the authentication and the *secure session* [2]. One of the important security features is that *Android 4.0* allows encrypted storage and remote data deletion [2]. This feature is the most helpful when the device is lost. The last feature of *Android 4.0* is the enhancement for enterprise (*VPN client API*). This feature allows us to construct the application to configure the addresses and routing rules, process outgoing and incoming packets and establish secure tunnels to remote server. This feature allows the application to be configured with *centric networks*. The main point of the implementation of this application is the usage of *Web Services*, which is going to provide better performance and possibility for calculation of the salaries for institutions having thousands of employees. By using the Cloud connection possibility, the application is going to result in saving energy, time and money of the institutions and numerous users stated in the conclusion and future work concepts of the paper.

II. RELATED WORK

Nowadays, the *Android* market is full of applications giving the users greater choice, but not all of them provide excellent quality to the user. There are many applications that have great marketing characteristics, but poor performances in terms of usability and efficiency. On the Android market several applications for calculating the salary can be found:

- Salary Bot by CAB Designs [10]
- Paycheck by Green App Developer [11]
- Quick Wage by CWE Software LLC [12]

Testing and analyzing these applications, *Salary Bot* [10] has the best performances among three of them, which allows the results to be broken yearly, monthly, weekly, daily, and hourly data; so, the design is much better than the other two applications. The common feature of these applications is the calculation of the gross salary with the additional work of the user. The *Paycheck* application [11] has simple and old design which is not consistent with the updates of Android and it provides approximate calculation of the salary. *Quick Wage* [12] allows the user to calculate the salary according to the wager. It also has old design which is not consistent with the Android updates and provides yearly, monthly, weekly, daily, and hourly salary calculation, providing salary calculation for several years. The disadvantage of *Quick Wage* is that it has an approximate calculation of salary.

The applications which are already developed allow the user to calculate the salary according to the user's gross earnings, but the gross earnings must also be calculated. The application that we decided to develop allows Web integration which gives the user easier way to calculate the salary. The user does not need to have previous knowledge on how to calculate the salary. This application is mainly developed for calculating the salary of the University staff, but it can be easily adjusted to other institutions as well. It is user-friendly, web integrated mobile application and its structure is presented in detail in the next section.

III. ANDROID APPLICATION

UniSal Android application is created to help the University staff mostly to calculate the salary in a user-friendly and efficient manner. The necessary information for calculating the salary is downloaded from the *Web Server* and parsed into usable information within the application (Figure 1). The information for calculating the salary is stored into *MySQL* database [14], which is stored into the *Web Server*. For the future development this application is going to give the *Cloud experience* to the user, so the application will be used by thousands of employees.

The application is calculating the gross salary based on monthly data for each employee in the institution, as well as on the basis of the prescribed rates and monthly average salary in our country.

The variables for calculating the Gross Salary *(GS)* are given as follows: Net Salary *(NS)*, Personal Income Tax *(PIT)*, and Contributions *(C)*. The equation for calculating gross salary is [6]:

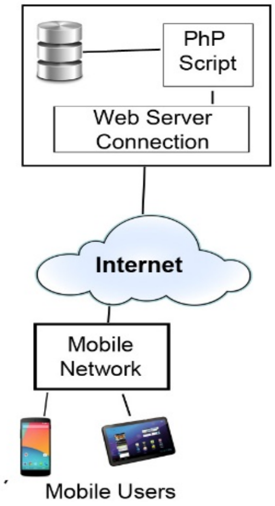


Figure 1. System architecture for mobile computing using Web Services.

$$GS = NS + PIT + C \qquad (1)$$

The category of salary is determined according to the law, the collective agreement and the employment contract. In order to calculate the salary, it is necessary to submit the data for its calculation to the salary referent in charge of data entry on the part of the authorized departments and managers in the firm. For calculating the salary the institution also has to calculate the personal income tax and the contributions. To calculate the amount of personal income tax and salary contributions starting the 1st of January *2009*, the rate of the personal income tax in the taxation of salaries is *10%* regardless of the amount of the employee's salary. Thereby the basis for the calculation of the personal income is the basis for the calculation of contributions (gross salary), less the total amount of the contributions and the personal exemption determined in monthly amount [6]. The variables for calculating the Basis for Personal Income Tax (BPIT) are: Gross Salary (GS), Contributions (C) and Personal Exemption (PE).

$$BPIT = GS – C – PE \qquad (2)$$

Contributions are calculated according to the rates determined into the Law regulations of the country. Table I describes the rates for the contributions [4][5]. The total sum of contributions for *2013* is described as (TSC).

Calculating contributions for the salary are presented in Table 1, as given below:

TABLE I. CONTRIBUTIONS FOR CALCULATING THE SALARY

Contributions	*Period*
	From the 1st of January
Mandatory PDI	18%
Mandatory health insurance	7.3%
Mandatory insurance in case of unemployment	1.2%
Additional contribution for compulsory health insurance in case of injury at work and occupational disease	0.5%

$$GS = NS + ((NS – PE) \times 11.111111) / 0.73 \qquad (3)$$

$$*0.73 = (100 – TSC) / 100$$

Equation (3) represents the way of calculating the GS of the employer [7].

A. Structure of the Application UniSal

This Android application *UniSal* is connected with remote database (*MySQL*) [14] and all required information for salary calculation is stored into the database. The information retrieved from the database is *JSON* [15] code, which is parsed when it is received into the application. Before it is parsed, the *JSON* code is received as a *JSON* Array and then each Object is accessed separately. All the values that are received from the database are stored into the Android device and are used for offline calculation of the salary. The *MySQL* database is saved into *Apache 2.0 Web Server* with *PhP 5* [16] *installation*. The *PhP scripts* are very important for this application, because the Android application does not need to have authentication information with the *MySQL* database. This provides *secure communication* between the *Android* application and *MySQL* database because all the information for authentication with the *MySQL* database is stored into the *PhP* scripts which can be accessed only from the server. When the Android application makes a *Http Post* request to the server, the server executes a specific script for returning the *JSON* code to the Android application.

The Main screen of the application is consisting of *ViewPager* that holds the fragments which are managed with *Section Adapter* and *Action Bar* that holds the tabs i.e. the style of the *Action Bar* is set to “Tab Style”. On the main screen, two tabs are displayed and they represent the calculation of the Gross and the Net Salary. The user interface is user-friendly and the navigation is done with swiping right-to-left or reverse.

The specific salary is calculated from the mandatory *PDI*, health insurance, contribution for employment, professional sick leave, total contributions, total contributions and charges. Another important section is the *Settings Panel* of the application. The settings panel allows the user to change the values of the contributions according to the changes done according to the Law regulations.

Once the application is launched, it checks the contributions on the remote server. If the *Android* device is connected on the *Internet*, the values for the contributions are downloaded automatically and stored into the device for offline calculation of the salary. The values are downloaded by sending a *Http Post* request to the server. The remote server holds *PhP* script for sending the values from *MySQL* database to the Android device. The script authenticates with the *MySQL* library and queries the values. Once the values are queried, they are parsed into *JSON* code and the code is sent to the *Android* device. When the *Android* device receives the *JSON* code it parses it and the values are saved into *xml* file for offline calculation of the salary. The Settings Panel is also available and allows the user to change the contribution values manually.

Due to improper changes into the Settings Panel, the application is not going to calculate the proper value of the salary for the employer. For this case scenario, the application generates error log holding the error values for the contributions. The error log is displayed to the user when he/she wants to calculate the salary. When the user presses the “Calculate” button the error log is displayed. The error log holds the exact contribution value which has error i.e. a value which is not reasonable for calculating the salary of the employee.

B. University staff salary calculation

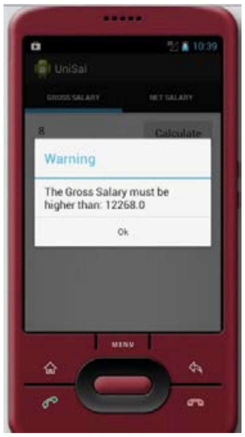
If the user wishes to calculate the salary based on points, he/she has to open the options menu on the Android device and tap on “Salary on points”. The application is going to open new activity containing the proper information for calculating the salary. The salary on points contains the following *Android* components: *Spinner*, *Button*, and *Checkbox*. The *Spinner* is used to hold the academic status; the *Checkbox* is used to determine whether the employee has been on a sick leave; and the *Button* is used to calculate the salary. The information into the *Spinner* is populated from *Remote Web Server*. When the information is downloaded into the device, it is stored into local database for offline calculation of the salary.

The information from the *Web Server* is retrieved by sending a *Http Post* request to the *Web Server*. Into the web server a *PhP* script is stored for retrieving the information to the *Android* device. The *PhP* script authenticates with the database on the server and queries the information that has to be sent to the Android device. The queried information is parsed into JSON code and sent to the Android device. On the Android hand side the information received is parsed into JSON Array. The information from the JSON Array is processed and it is saved into the local *SQLite* database for offline calculation of the salary. When the user presses the "Calculate" button, the salary is calculated according to the information downloaded from the remote server. If the "Sick leave" option is marked and the "Calculate" button is pressed, the calculated salary amount is going to equal *70 %* of the full amount. This is determined by the Law regulations in the country.

IV. TESTING THE APPLICATION AND RESULTS

The application testing is performed in a couple of steps: testing the Gross and Net salary, testing the Settings Panel and testing of University staff salary.

A. *Testing the gross and the net salary*



The Gross Salary must be higher than: 12268.0

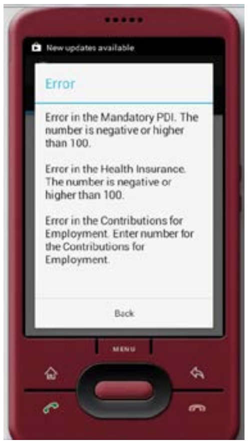
Figure 2. Gross Salary test result.

For the gross and the net salary, we have applied the test scenarios - how the application responds when there is no active Internet connection, how the application responds when *0* is entered as a value and the last test is how the application responds when the negative button is returned as a value. For the first test scenario we have entered gross salary which is below the minimum gross salary determined by the Law regulations. Figure 2 gives the result when the application is notifying the user that he/she needs to enter salary greater than the minimum salary - the result is the same when the user enters *0*. Of course, the user can enter only positive numbers. In the next test, we have left the field for the salary empty and tried to calculate the salary. When the "Calculate" button was pressed, the application responded with a message that the user has to enter salary. If the user is trying to download the data from the Web Server when there is no Internet connection, the user is going to receive notification that there is problem with the Internet connection. The same test scenarios were applied for the net salary. The difference between these two types of salaries is the minimum salary determined according to the Law regulations. The response of the application when the salary amount is below the minimum salary amount of the country is also presented in Figure 2.

B. *Testing the Settings Panel*

This test is used to determine whether the user has made any mistakes inserting the values for the contributions when there is no internet connection. The information for the contributions which is downloaded from the *Web Server* is updated according to the Law regulations and the application is updating accordingly. When the user is unable to connect to the Internet and there is a need for calculating the salary urgently, the user can enter the values for the contributions manually. When the application is performed by the user, the latest values for the salary contributions are already inserted.

This case scenario is displaying the result when the user has made a mistake entering the contribution values. Figure 3 presents the result when the user has made a mistake entering the values manually. This type of error dialog is only generated if the user has entered a value below zero, a value above one hundred and instead of number has entered an alphabetic character. If the value is below zero, it means that the government should pay the user, which is not real; if the value is above one hundred that means that the whole salary and more should go for the government, which is also not real. When a character is entered as a value, the error dialog is also generated, because a positive number is required as a value..



Error in the Mandatory PDI. The number is negative or higher than 100.

Error in the Health Insurance. The number is negative or higher than 100.

Error in the Contributions for Employment. Enter number for the Contributions for Employment.

Figure 3. Settings Panel test result.

C. *Testing the University staff salary*

This test is performed to determine whether the values are successfully downloaded and correctly displayed into the application. When the user enters the application section for calculating the University staff salary, the values are automatically downloaded if there is available internet connection. If there is no Internet connection, the user is notified that he needs to connect to the internet in order to calculate the salary. If the "Calculate" button is pressed when the values are not downloaded the user is going to receive notification that the value cannot be zero. If no value is returned the application is considering that the value is zero. If the application is connected to the Internet the values are downloaded, parsed and added into the spinner. Now, the user can calculate the salary according to the downloaded values. If the checkbox "Sick leave" is marked, *70 %* of the salary is calculated according to the Law regulations.

V. QUALITY OF EXPERIENCE (QOE) OF MOBILE COMPUTING

Quality of Service (QoS) refers to the technical operational aspects of a service, such as time to support services, capacity, and transport. *Quality of Experience (QoE)* measures the difference between what users expected and what they actually received. Using the *QoE* is beneficial to estimate the perception of the user about the quality of a particular service and it depends on the customer's satisfaction in terms of usability, accessibility, retaining ability and integrity of using specific service [13]. *QoE* means overall acceptability of an application or service, as perceived subjectively by the end-user and represents multidimensional subjective concept that is not easy to evaluate (Figure 4). In our work, we have used *QoE* evaluation in order to measure the quality of the mobile application usage.

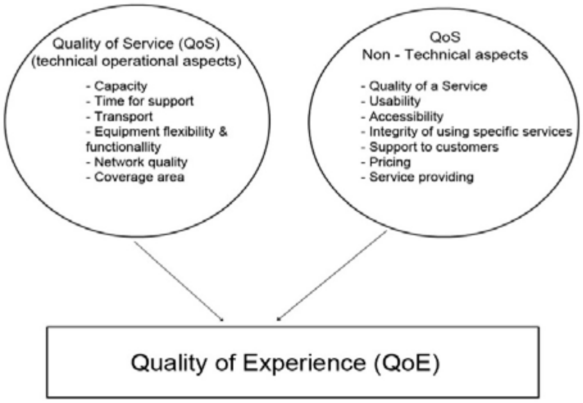


Figure 4. Relation between QoS & QoE.

The results presented in the following chart were obtained according the Quality of Experience evaluation performed with the University staff. The application is tested and evaluated by different scenarios. The survey questions were answered by a group of representatives of the University staff that participated in the *Android* application implementation locally (*UniSal*) and using Web Services (*UniSal Web*) as well. Analyzing the answers from the Mobile application implementation using *Web Services* has provided with the summary given in histogram presented in Figure 5.

The first concept evaluation was the quality of the *Graphical User Interface* of the *Android* application. By the observation the following things can be concluded: *76 %* of the University staff participants concluded that the graphical design of the Web application was *Android* consistent and easy to use; they were supportive about the design and suggest the idea for the colors to be according to the University colors.

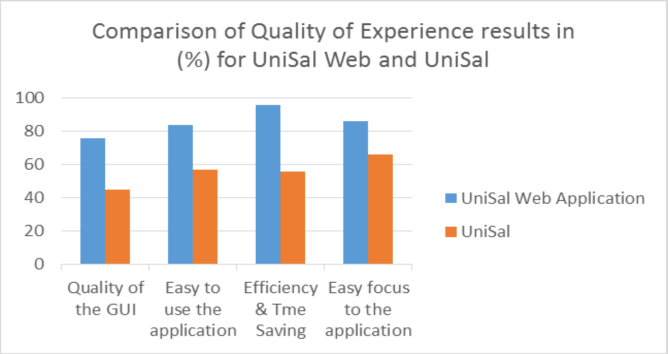


Figure 5. Comparison of QoE results in %.

The second evaluation concept was Usability of the Android application. As presented in Figure 5, *84 %* of the University staff concluded that the application is very efficient and usable because it is *Web* based application and all necessary information for calculating the salary is downloaded automatically.

The third evaluation scenario was concerned with the Efficiency and Time Saving. *96 %* of the University staff concluded that by using this application time is saved when calculating the salary, because all the information is downloaded from the Internet, so energy, time, and money are saved. It is an efficient mobile application, while 56% were declaring the efficiency of *UniSal Android application* which is not *Web* oriented.

86 % of the University staff concluded that it is easy to focus to the Web-oriented Android application. The other 66% were focusing on the *UniSal*, local server supported application.

Most of the users were satisfied with the Web-oriented mobile application, its usability, efficiency and availability referring to the *Web Services* users and of course, *Cloud connection possibility* providing to the application.

VI. CONCLUSION

The *Android UniSal Web application* is divided into two sections. These are calculating the Gross and the Net Salary for the University staff. The application has the functionality that allows the user to update the contributions if they have been changed according to the new Law regulations. The user can change the following contributions: *PDI, health insurance, contributions, personal tax and personal income*. If the user makes mistake inserting the contributions, the application generates error log so that the user can fix the contributions. The application allows the user of to check whether the employee had sick leave in order to calculate the salary. To ensure that the application is working according to the inserted parameters the application has passed couple of tests scenarios - tests for *University staff* gross and net salary calculation, tests for the Setting Panel and the menu of contributions, entering unreal numbers in order to check whether the application will accept the inputs or it will generate error log to the user.

The *mobile web application* presented in this work proposed an original, efficient way of managing the documentation in an institution. Our approach improves the *Quality of Experience (QoE)*; provides the users all necessary resources and performances, which refers to a user-friendly mobile application; web services oriented, evaluated by *QoE*, referring the quality of the *graphical user interface*, *usability*, *efficiency and time saving*.

By using the *mobile cloud computing connection possibility* of this *Android* application, as a future work, more advantages would be provided, such as *flexibility, portability and scalability* that where obstacles of the mobile devices so far. By observing the *QoE* survey, the advantages from using this application have significantly increased users' attention. Possibility of *mobile computing* technology has provided improvement in the process of *Android Web application* development in the direction of increasing the quality of services.

The application presented in this work is developed by using the graphical components provided from the Android API. The key graphical components for this application are the Action Bar and the Fragment implementation. The user does not need to have a special background for salary calculation i.e. the salary is calculated simply by clicking one button. All the necessary information required for calculating the salary is downloaded automatically from the Web Server. In time, the application would require hardware with higher performance, and other APIs to work faster and to process the information. By using the cloud connection, this would not be a problem since the Cloud provider allows using different plans Virtual Machines and APIs for the application. As part of our future work, we are developing an application with Cloud connection. The “Brain Application” which is handling and processing the request is Java EE application and it will be uploaded into the Cloud. By using this model and infrastructure it allows us to enlarge our work in different technologies and creating more complex and more useful options to the user, such as iOS integration, Windows Mobile integration, creating suitable desktop application. Some of the additional options are the following: Employee Management (adding, removing and updating employees), Salary Management (creating, removing and updating the salaries), Salary Calculation for certain employee.

We believe that these Web-oriented configurations of mobile applications will be especially beneficial for universities and their academic and administrative staff, by promoting the advantages from using such an application in terms of saving energy, time, and money.

VII. ADVANTAGES AND FUTURE WORK

The *Android UniSal Web application* could be improved in terms of using *Android 4.0,* which has new features that are still upgrading and new features are coming, such as *NFC* system, which allows the smartphones’ users to unlock cars, pay for parking bills, pay the food into a restaurant and the most common usage is transferring images by just touching the smartphone. With this functionality, this application in the future can be used as a check-in card into some institutions or to calculate the salary by just touching the phone of the employer and the phone of the employee. By using the cloud services in the future, and the proper security issues that *Android 4.0* delivers, the information of the employees can be stored on *cloud database* and the employer can analyze the work of his employees from his phone while he is on his way to work. An interesting *Security* concept for this application given by the unlocking the app with face recognition unlock system that is integrated into *Android 4.0* and higher. The employer who is using the app can set a security lock from his face and he will be the only person who can load the application, calculate the salary or analyze the work of the employee and much more. This mobile application will be easily considered as an adjustable content for *Content Centric Network (CCN)* and *Information Centric Networks (ICN)* approach, aiming to achieve efficient and reliable distribution of the content by providing general platform for communication services which refers to a part of our future work.

REFERENCES

[1] D. Morril, “Announcing the Android 1.0 SDK,” Release 1, Android Developers Blog, 2008, http://android-developers.blogspot.com/2008/09/announcing-android-10-sdk-release-1.html, [Retrieved: May, 2014]

[2] Ice Cream Sandwich (n.d.) In Developers, http://developer.android.com/about/versions/android-4.0-highlights.html, [Retrieved: May, 2014]

[3] P. Alto, “Google’s Android becomes the world’s leading smart phone platform,” Canalys, 2011, http://www.canalys.com/newsroom/google%E2%80%99s- android-becomes-world%E2%80%99s-leading-smart-phone-platform, [Retrieved: May, 2014]

[4] Health Insurance Fund, “Instructions for the manner of calculating contributions for compulsory Health Insurance, according to the amendments of the Health Insurance Law,” 2007.

[5] Ministry of Finance, “Instructions for the implementation of the concept for calculation and payment of the gross salary of the employees in the institutions - budget beneficiaries,” 2009.

[6] Public Revenue Office of the Republic of Macedonia, “Gross salary,” 2013.

[7] Public Revenue Office of the Republic of Macedonia, “Calculation of salary,” 2013.

[8] G. Allen, “Beginning Android 4 Apress,” 2012, http://www.apress.com/9781430239840, [Retrieved: May, 2014]

[9] E. Burnette, “Hello, Android: Introducing Google's Mobile Development Platform (Pragmatic Programmers),” Pragmatic Bookshelf: Third Edition, Released: August 4, 2010.

[10] Salary Bot (n.d.) https://play.google.com/store/apps/details?id=salaryCalculator.salaryCalc, [Retrieved: May, 2014]

[11] PayCheck (n.d.) https://play.google.com/store/apps/details?id=salary.calculat, [Retrieved: May, 2014]

[12] Quick Wage (n.d.) https://play.google.com/store/apps/details?id=com.cwesoftware.salary, [Retrieved: May, 2014]

[13] W. Wu, A. Arefin, R. Rivas, K. Nahrstedt, R. M. Sheppard, and Z. Yang, “Quality of Experience in Distributed Interactive Multimedia Environments: Toward a Theoretical Framework,” MM '09 Proceedings of the 17th ACM international conference on Multimedia, October, 2009, pp. 481-490, DOI: 10.1145/1631272.1631338.

[14] MySQL Official Site (n.d.) Retrieved from http://www.mysql.com/, [Retrieved: May, 2014]

[15] JSON Official Site (n.d.) Retrieved from http://json.org/, [Retrieved: May, 2014]

[16] PhP Official Site (n.d.) Retrieved from http://www.php.net/, [Retrieved: May, 2014]