AICT 2012

The Eighth Advanced International Conference on Telecommunications

ISBN: 978-1-61208-199-1

May 27- June 1, 2012

Stuttgart, Germany

AICT 2012 Editors

Michael Massoth, University of Applied Sciences - Darmstadt, Germany

Michael D. Logothetis, University of Patras, Greece

Dragana Krstic, University of Nis, Serbia

AICT 2012

Forward

The Eighth Advanced International Conference on Telecommunications (AICT 2012) held on May 27 - June 1, 2012 - Stuttgart, Germany, 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.

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.

E-learning refers to on-line learning delivered over the World Wide Web via the public Internet or the private, corporate intranet. The conference considered how, when and where e-learning helps to solve the training needs, what the challenges of creating and managing vast amounts of e-learning are, how the upcoming IT technologies influence e-learning and how the Web based educational materials should be developed to meet the demands of the long-life, motivated and very often self-directed students.

The conference also addressed teletraffic modeling and management. It covered traffic theory, traffic control and QoS, performance evaluation methods, network design and optimization of wired and wireless networks, and simulation methodology for communication networks.

We take this opportunity to thank all the members of the AICT 2012 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 the AICT 2012. We truly believe that, thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the AICT 2012 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that AICT 2012 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in telecommunications.

We are convinced that the participants found the event useful and communications very open. The beautiful city of Stuttgart surely provided a pleasant environment during the conference and we hope you had a chance to visit the surroundings.

AICT 2012 Chairs
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

AICT Special Area Chairs
TELET
Mariusz Glabowski, Poznan University of Technology, Poland
Denis Collange, Orange Labs - Sophia Antipolis, France

Optical
Djafar K. Mynbaev, New York City College of Technology - Brooklyn, USA

AICT 2012

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

AICT Special Area Chairs

TELET
Mariusz Glabowski, Poznan University of Technology, Poland
Denis Collange, Orange Labs - Sophia Antipolis, France

Optical
Djafar K. Mynbaev, New York City College of Technology - Brooklyn, USA

AICT 2012 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
Pedro A. Aranda Gutiérrez, University of Paderborn, Germany
Loredana Arienzo, Joint Research Centre - European Commission - Ispra, Italy
Miguel Arjona Ramírez, University of São Paulo, Brazil
Andres Arjona, Nokia Siemens Networks, Japan
Michael Atighetchi, Raytheon BBN Technologies-Cambridge, USA
Tulin Atmaca, TELECOM SudParis, France
Marco Aurélio Spohn, Federal University of Campina Grande, Brazil
Konstantin Avratchenkov, INRIA- Sophia Antipolis, France
Paolo Barsocchi, ISTI/National Research Council - Pisa, Italy
Ilija Basicevic, University of Novi Sad, Serbia
Carlos Becker Westphall, Federal University of Santa Catarina, Brazil
Daniel Benevides da Costa, Federal University of Ceará (UFC), Brazil
Ilham Benyahia, Université du Québec en Outaouais, Canada
Eugen Borcoci, University "Politehnica" of Bucharest (UPB), Romania
Christos Bouras, University of Patras, Greece
Julien Broisin, Université Paul Sabatier, Toulouse III, France
Prasad Calyam, The Ohio State University, USA

Maria-Dolores Cano Banos, Universidad Politécnica de Cartagena, Spain
Fernando Cerdan, Universidad Politecnica de Cartagena, Spain
Tijani Chahed, Telecom SudParis, France
Hakima Chaouchi, Telecom SudParis, France
Phool Singh Chauhan, Indian Institute of Technology Kanpur, India
Lee Feng Cheng, Hewlett-Packard, Singapore
Stefano Chessa, University of Pisa, Italy
Sungsoo Choi, Korea Electrotechnology Research Institute (KERI), S. Korea
Denis Collange, Orange Labs - Sophia Antipolis, France
Todor Cooklev, Indiana-Purdue University - Fort Wayne, USA
Carlton Davis, École Polytechnique de Montréal, Canada
Chérif Diallo, Consultant Sécurité des Systèmes d'Information, France
Quang Trung Duong, Blekinge Institute of Technology, Sweden
Zbigniew Dziong, École de Technologie Supérieure - Montreal, Canada
Mohamed El-Tarhuni, American University of Sharjah , UAE
Mario Fanelli, University of Bologna, Italy
Mário F. S. Ferreira, University of Aveiro, Portugal
Bruno Filipe Marques, Polytechnic Institute of Viseu, Portugal
Pedro Fortuna, University of Porto, Porutgal
Paraskevi Fragopoulou, TEI of Crete, Greece
Alex Galis, University College London, UK
Christos K. Georgiadis, University of Macedonia - Thessaloniki, Greece
Marc Gilg, University of Haute Alsace, France
Mariusz Glabowski, Poznan University of Technology, Poland
Katie Goeman, Hogeschool-Universiteit Brussel, Belgium
Stefanos Gritzalis, University of the Aegean, Greece
Vic Grout, Glyndwr University - Wrexham, UK
Lei Guo, Northeastern University, China
Ibrahim Habib, City University of New York, USA
Go Hasegawa, Osaka University, Japan
Michiaki Hayashi, KDDI R&D Laboratories Inc., Japan
Mannaert Herwig, University of Antwerp, Belgium
Toan Hoang, Norwegian Defense Research Establishment, Norway
Ilias Iliadis, IBM Zurich Research Laboratory, Switzerland
Muhammad Ali Imran, University of Surrey - Guildford, UK
Lucian Ioan, University: "Politehnica" of Bucharest (UPB), Romania
Michail Kalogiannakis, University of Crete, Greece
Giorgios Kambourakis, University of the Aegean - Karlovasi, Greece
Charalampos Karagiannidis, University of Thessaly - Volos, Greece
Ziad Khalaf, SUPELEC/SCEE, France
Kashif Kifayat, Liverpool John Moores University, UK
Insoo Koo, University of Ulsan, Korea
Francine Krief, Université de Bordeaux - IPB, France
Robert Koch, University of the Federal Armed Forces / German Navy, Germany
Dragana Krstic, University of Nis, Serbia
Thomas D. Lagkas, University of Western Macedonia - Thessaloniki, Greece

Brian Lee, Software Research Institute, Ireland
Keqin Li, State University of New York - New Paltz, USA
Jia-Chin Lin, National Central University, Taiwan, ROC
Diogo Lobato Acatauassú Nunes, Federal University of Pará - Belém, Brazil
Michael D. Logothetis, University of Patras, Greece
Renata Lopes Rosa, University of São Paulo, Brazil
Zoubir Mammeri, IRIT - Toulouse, France
Michel Marot, Telecom SudParis, France
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 - Tripolis Greece
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
Nokolai Nefedov, Nokia Research Center, Switzerland
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
Harald Øverby, Norwegian University of Science and Technology - Gløshaugen, Norway
Constantin Paleologu, University Politehnica of Bucharest, Romania
Jari Palomäki, Tampere University of Technology - Pori, Finland
Andreas Papazois, RACTI & CEID / University of Patras, Greece
Cathryn Peoples, University of Ulster, UK
Fernando Pereñíguez García, University of 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
Andreas Pitsillides, University of Cyprus-Nicosia, Cyprus
Adrian Popescu, Blekinge Institute of Technology - Karlskrona, Sweden
Dusan Radovic, TES Electronic Solutions GmbH - Stuttgart, Germany
Ustijana Rechkoska Shikoska, University for Information Science & Technology "St. Paul the Apostle" - Ohrid, Republic of Macedonia
Eric Renault, Telecom SudParis, France
Lorayne Robertson, University of Ontario Institute of Technology, Canada
Danguole Rutkauskiene, Kaunas University of Technology, Lithuania

Demetrios G. Sampson, University of Piraeus & CERTH, Greece
Panagiotis Sarigiannidis, University of Western Macedonia - Kozani, Greece
Reijo Savola, VTT Technical Research Centre of Finland - Oulu, Finland
Sergei Semenov, Nokia Corporation, Finland
Sandra Sendra Compte, University Polytechnic of Valencia, Spain
Michelle Sibilla, Paul Sabatier University Toulouse 3, France
Nicolas Sklavos, Technological Educational Institute of Patras, Hellas
Keattisak Sripimanwat, National Science and Technology Development Agency (NSTDA), Thailand
Lars Strand, Nofas Management, Norway
Daniele Tafani, Dublin City University, Ireland
Yutaka Takahashi, Kyoto University, Japan
Tomohiko Taniguchi, Fujitsu Laboratories Limited, Japan
Yoshiaki Taniguchi, Osaka University, Japan
Richard Trefler, University of Waterloo, Canada
Thrasyvoulos Tsiatsos, Aristotle University of Thessaloniki, Greece
Kazuya Tsukamoto, Kyushu Institute of Technology-Fukuoka, Japan
Kenneth Turner, The University of Stirling, Scotland
Masahiro Umehira, Ibaraki University, Japan
Guillaume Valadon, French Network and Information Security Agency, France
John Vardakas, University of Patras, Greece
Manos Varvarigos, University of Patras, Greece
Dimitris Vasiliadis, University of Peloponnese Greece
Calin Vladeanu, University Politehnica of Bucharest, Romania
Benno Volk, E-Learning Center (ELC) / University of Zurich, Switzerland
Luca Vollero, Università Campus Bio-Medico di Roma, Italy
Mea Wang, University of Calgary, Canada
Amali Weerasinghe, University of Canterbury, New Zealand
Steve Wheeler, University of Plymouth, UK
Bernd E. Wolfinger, University of Hamburg, Germany
Mudasser F. Wyne, National University - San Diego, USA
Kang Xi, Polytechnic Institute of New York University, USA
Qin Xin, Université Catholique de Louvain - Louvain-la-Neuve, Belgium
Miki Yamamoto, Kansai University, Japan
Qing Yang, Ciena Corporation, USA
Vladimir S. Zaborovsky, Technical University - Saint-Petersburg, Russia
Giannis Zaoudis, University of Patras, Greece
Demóstenes Zegarra Rodríguez, University of São Paulo, Brazil
Liaoyuan Zeng, University of Limerick, Republic of Ireland
Rong Zhao, Detecon International GmbH - Bonn, Germany
Zuqing Zhu, University of Science and Technology of China, China
Martin Zimmermann, Hochschule Offenburg - Gengenbach, 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

A Two-Stage Pilot-Assisted CFO Estimation Scheme for OFDM Signals

Omar Al-Sammarraie
Department of Electrical Engineering
American University of Sharjah
Sharjah, UAE
b00024985@aus.edu

Mohamed El-Tarhuni
Department of Electrical Engineering
American University of Sharjah
Sharjah, UAE
mtarhuni@aus.edu

***Abstract*— In this paper, a pilot-assisted two-stage carrier frequency offset (CFO) estimation algorithm is proposed for OFDM signals. The proposed scheme performs a coarse search over a window of possible carrier offsets to obtain an initial estimate of the CFO. Then, a fine search is carried over a much smaller search window to obtain better accuracy of the carrier frequency offset. The pilot signal used in this algorithm results in a slowly varying correlation function with peaks at the correct offset. This is exploited by the proposed algorithm to reduce the search window for the second stage. Simulation results presented in this paper show that a significant improvement in the mean-square error performance is achieved by using this two-stage approach without increasing the complexity of implementation compared to single-stage conventional algorithms. It is also demonstrated that the proposed scheme outperforms the conventional cyclic-prefix based CFO estimation algorithm under both flat and frequency-selective Rayleigh fading channel conditions.**

Keywords-Carrier Frequency Offset Estimation, OFDM Synchronization; Multi-stage Search.

I. INTRODUCTION

Orthogonal Frequency Division Multiplexing (OFDM) has been proposed to support high data rate applications in future mobile radio systems. This is due to its improved spectral efficiency and efficient implementation structure. Furthermore, OFDM has inherent features in terms of mitigating multipath and intersymbol interference (ISI) that are dominant in high data rate scenarios [1]–[3]. OFDM avoids the ISI problem by transmitting the data over a large number of narrow band channels (subcarriers) and by using a cyclic prefix at the start of every OFDM symbol.

A successful deployment of OFDM-based systems, however, needs to overcome several challenges. One of the main problems is the high peak-to-average power ratio of the transmitted signal resulting in a loss of power efficiency. Another major issue, which is the focus of this paper, is the need for accurate subcarrier synchronization, also called carrier frequency offset (CFO) estimation, to maintain orthogonality between the subcarriers and, consequently, prevent inter-carrier interference (ICI). Frequency offsets occur in OFDM systems mainly due to frequency mismatches between the transmitter and receiver oscillators and partly due to Doppler shifts owing to user mobility. Since all the subcarriers should be orthogonal to one another to ensure successful data recovery, frequency offsets pose a major issue that must be accurately resolved for a successful OFDM system implementation [4][5].

OFDM frequency offset estimators can be classified into non pilot-based and pilot-based techniques. Non pilot-based schemes rely on the OFDM symbol structure features such as the cyclic prefix (CP) [6]. These schemes are more bandwidth efficient than pilot-based schemes but they tend to suffer under frequency-selective fading channel conditions [7]. However, pilot-based schemes tend to be more accurate in estimating the CFO. Furthermore, the use of pilot-based schemes is justified since the pilot data can be utilized for other purposes such as channel estimation.

Pilot-based CFO estimation schemes use known training symbols sent by the transmitter over specific subcarriers and time slots. The pilot symbols may be transmitted as a preamble over all subcarriers during the first OFDM symbol and then regular data transmission is started. Different maximum likelihood (ML) CFO estimators were developed in [8]-[13] for both flat and frequency-selective fading channels. Although ML schemes provide good performance in terms of the mean square estimation error but they tend to require high computational complexity.

In this paper, we introduce a two-stage pilot-based CFO synchronization technique that provides accurate carrier offset estimation. The proposed scheme uses different search steps to find the offset resulting is a reduced complexity compared to single-stage searching strategy. The performance of the proposed scheme is presented under both flat and frequency selective fading channel conditions.

The rest of the paper is organized as follows: Section II describes the system model. The proposed scheme is presented in Section III. Simulation results and conclusions are given in Sections IV and V, respectively

II. SYSTEM MODEL

The block diagram of the OFDM system under consideration in this work is shown in Figure 1. An OFDM signal is formed by modulating several orthogonal subcarriers with different data symbols. The input binary data is first applied to the modulation block that maps the input data to corresponding modulation symbols. Then, serial to parallel (S/P) conversion is performed followed by inverse Fast Fourier Transform (IFFT) operation to distribute the symbols to the different subcarriers. A cyclic prefix is added to form an OFDM symbol and then converted from parallel

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

to serial (P/S) for transmission after digital to analog conversion (DAC).

The transmitted OFDM symbol is expressed as

$$x(t) = \sum_{k=0}^{N-1} b_k \, e^{j2\pi kt/T}, \quad (1)$$

where T is the symbol period, N is the number of orthogonal subcarriers, and b_k is the data symbol of the kth subcarrier. A sampled version of the OFDM signal is obtained by taking samples every nT/N seconds, where n is an integer, to get

$$x(n) = \sum_{k=0}^{N-1} b_k \, e^{\frac{j2\pi kn}{N}}. \quad (2)$$

The OFDM signal is transmitted through a wireless channel and experiences fading, noise, and carrier frequency offset to obtain the received signal given by

$$y(t) = e^{j2\pi\Delta ft} \{h(t) * x(t)\} + \omega(t), \quad (3)$$

where $h(t)$ is the impulse response of the channel, $\omega(t)$ is the additive white Gaussian noise (AWGN), and Δf is the unknown frequency offset (in Hz) to be estimated. In discrete form, this reduces to

$$y(n) = e^{\frac{j2\pi\rho n}{N}} \{h(n) * x(n)\} + \omega(n); \quad n = 0,1, \ldots \quad (4)$$

where $\rho = \Delta fT$ is the normalized frequency offset (frequency offset normalized by the symbol rate). For a flat fading channel, the received signal is written as

$$y(n) = \alpha e^{\frac{j2\pi\rho n}{N}} \sum_{k=0}^{N-1} b_k \, e^{\frac{j2\pi k(n-\tau)}{N}} + \omega(n), \quad (5)$$

where α is the complex channel gain modeled with Rayleigh amplitude and uniform phase and τ is the channel delay. In case of slowly varying channels, this represents a sinusoidal signal shifted in frequency by ρ and multiplied by a constant complex number.

To recover the transmitted data, the receiver performs reverse operations on the received signal as shown in Fig. 1. However, a frequency synchronization block is needed to estimate the unknown frequency offset that is provided to the FFT block for correction. The proposed frequency synchronization algorithm is discussed in the following section.

III. PROPOSED SYSTEM

The proposed CFO estimation scheme, shown in Fig. 2, performs a search over a window of possible frequency offsets in order to estimate the unknown offset Δf. The scheme uses a two-stage, also known as double-dwell (DD), search strategy. The first stage performs a coarse search over the possible frequency offset search window with a relatively large frequency separation (called step size) between the test offsets in order to reduce the complexity of the synchronization algorithm. The second stage uses a much smaller step size than the first stage but searches over a much smaller window of possible offsets as specified by the largest correlation peaks from the first stage. This results in a significant improvement in the estimation accuracy while maintaining a low complexity of implementation.

We assume that a known pure sinusoidal preamble pilot data sequence, $\{b_k^P\}$, is transmitted for synchronization purposes. Although using a pilot sequence causes some degradation to the bandwidth efficiency, especially in fast fading channels, these pilots are needed for channel estimation and hence their use for accurate frequency offset estimation is justified. Furthermore, the distribution of pilot symbols over the time-frequency grid can vary depending on channel conditions (time and frequency selectivity). During the pilot sequence transmission, the received signal is reduced to

$$y(n) = \alpha \sum_{k=0}^{N-1} b_k^P e^{\frac{-j2\pi k\tau}{N}} e^{\frac{j2\pi(k+\rho)n}{N}} + \omega(n). \quad (6)$$

The first stage of correlation is performed by correlating the received signal in (6) with a number of complex-conjugate versions of the pilot sequence each with a test normalized frequency offset $\gamma_m = m\epsilon_1$, where $\epsilon_1 = ST$ is the normalized search step size for the first stage and S is the search step size in Hz and m is an integer with values $m = -\frac{1}{2\epsilon_1}, -\frac{1}{2\epsilon_1} + 1, \ldots -1, 0, 1, \ldots, \frac{1}{2\epsilon_1} - 1, \frac{1}{2\epsilon_1}$. Without loss of generality, we assume $\frac{1}{\epsilon_1}$ to be an even number so we have m as an integer. The step size for the first stage is typically large, e.g., 10% of the frequency offset $S = \Delta f/10$ in order to reduce the number of correlations and hence reduce the complexity. For the m^{th} offset, correlating the received signal with the known pilot symbols during the first search stage results in

$$R^{(1)}(m) = C \sum_{n=0}^{N-1} e^{\frac{j2\pi(\rho-\gamma_m)n}{N}} + v(m), \quad (7)$$

where $C = \alpha \sum_{k=0}^{N-1} |b_k^P|^2$ is a complex constant and $v(m)$ is due to the noise component.

The objective is to find the value of γ_m that maximizes the magnitude of the correlation function in (7). It can be shown that this is equivalent to maximizing the following function over γ_m

$$U(m) = \left\{ \frac{\left[\left(1-\cos\left(2\pi(\rho-\gamma_m)\right)\right)^2 + sin^2\left(2\pi(\rho-\gamma_m)\right)\right]}{\left[\left(1-\cos\left(\frac{2\pi(\rho-\gamma_m)}{N}\right)\right)^2 + sin^2\left(\frac{2\pi(\rho-\gamma_m)}{N}\right)\right]} \right\}, \quad (8)$$

which is maximized when $\gamma_m = \rho$; i.e., when the test offset equals to the actual frequency offset to be estimated. Fig. 3 displays some examples for the above function for a normalized frequency offset ρ of -0.3, 0, and 0.3, demonstrating that the maximum is achieved when the test offset equals to the actual frequency offset.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

In practical implementations, only a finite number of test offsets are searched. Thus, it is possible to have an error due to the limited resolution of the search process. This error can be reduced by increasing the number of search steps, i.e., reduce the step size. However, the hardware and computational power could be used more efficiently if the predictability of the correlation function is exploited since the exact offset corresponds to the function's global maximum. Therefore, if the exact offset does not coincide with the search steps, it must then be between the two offsets corresponding to the two highest correlation peaks.

Thus, we use a second search stage with a smaller step size, $\epsilon_2 \ll \epsilon_1$ in order to improve the accuracy of the estimation. The search window is limited to the range between the two frequency offsets with the largest correlation values obtained over the first search stage such that the computational complexity is reduced. Suppose that $|U(m)|^2$ has it is largest or peak value at $m = m_P$, then the second stage search is either performed between γ_{m_P} and γ_{m_P+1} if $|U(m_P+1)|^2 > |U(m_P-1)|^2$ or between γ_{m_P} and γ_{m_P-1} if $|U(m_P+1)|^2 < |U(m_P-1)|^2$. The search is done in steps of ϵ_2 and thus limits the maximum offset estimation error to $\pm\epsilon_2/2$.

The following correlation operation is performed for the second stage

$$R^{(2)}(l) = C\sum_{n=0}^{N-1} e^{\frac{j2\pi(\rho-\vartheta_l)n}{N}} + g(l), \qquad (9)$$

where ϑ_l is the normalized test offset and $g(l)$ is the noise component. The normalized test offset is determined as:

$$\vartheta_l = \begin{cases} \gamma_{m_P+l\epsilon_2} & \text{if } |U(m_P+1)|^2 > |U(m_P-1)|^2 \\ \gamma_{m_P-l\epsilon_2} & \text{if } |U(m_P+1)|^2 < |U(m_P-1)|^2 \end{cases}, \qquad (10)$$

for $l = 0,1,2,\ldots,\lceil 1/\epsilon_2 \rceil$ where $\lceil q \rceil$ is the integer value greater than or equal to q. Finally, the offset that results in the largest correlation in (9) is used as the estimate for the actual normalized frequency offset $\hat{\rho}$. The estimated offset is then provided to the FFT block at the receiver for offset correction prior to the FFT operation.

IV. Simulation Results

In this section, the performance of the proposed double-dwell (DD) synchronization technique is presented in terms of the mean-square error (MSE) in estimating the frequency offset. The proposed scheme performance is compared to the performance of the conventional cyclic extension (CE) synchronization technique. The channel is assumed to follow a Rayleigh fading model with either flat fading or frequency selective fading. The OFDM system is assumed to have N = 128, 256, or 512 subcarriers and the cyclic extension is 1/8th of the OFDM symbol duration. The double-dwell scheme uses a first stage with a normalized search step size of $\epsilon_1 = 0.1$ and a second stage with a normalized search step size of $\epsilon_2 = 0.01$. These values were used for illustration purposes and other values could be used without loss of generality.

Fig. 4 compares the MSE performance of the double-dwell scheme with that of the cyclic-extension based scheme under flat fading for an OFDM system with 128, 256, or 512 subcarriers. The results show that the double-dwell scheme provides a significant improvement, especially at low signal-to-noise ratio, with a gain of more than one order of magnitude. It is observed that the performance for both schemes improves as the number of subcarriers increases.

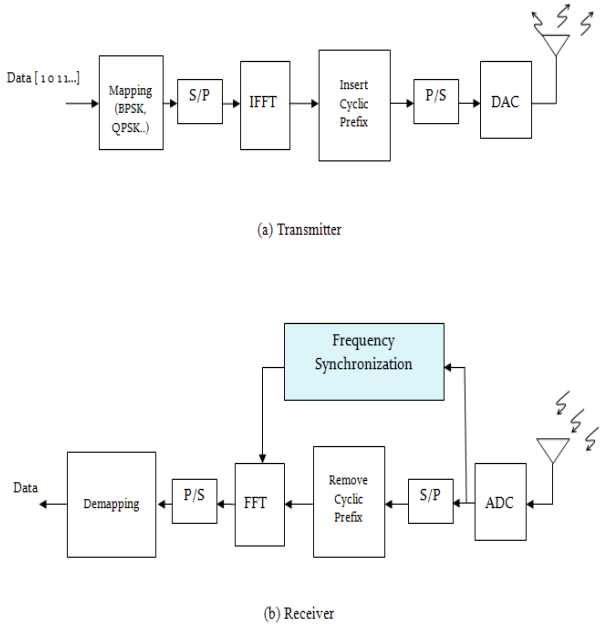


Figure 1. OFDM system block diagram.

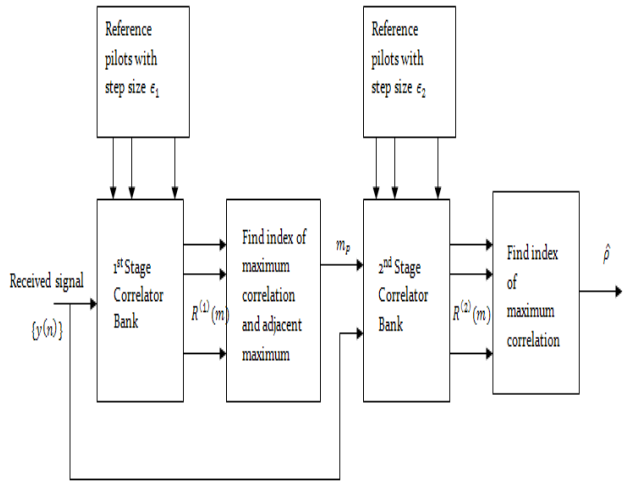


Figure 2. Block diagram of the proposed double-dwell frequency synchronization scheme.

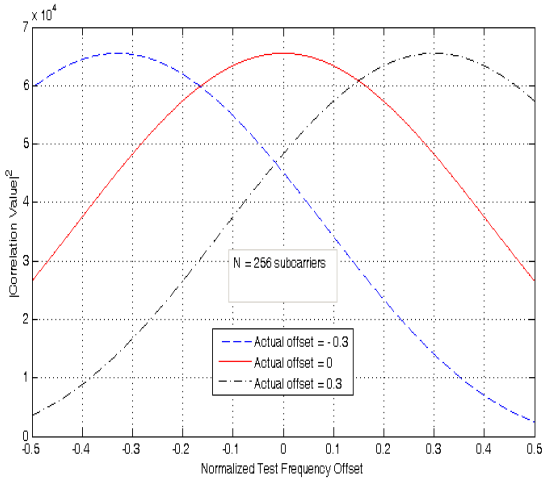


Figure 3. Examples of the correlation function for different frequency offsets.

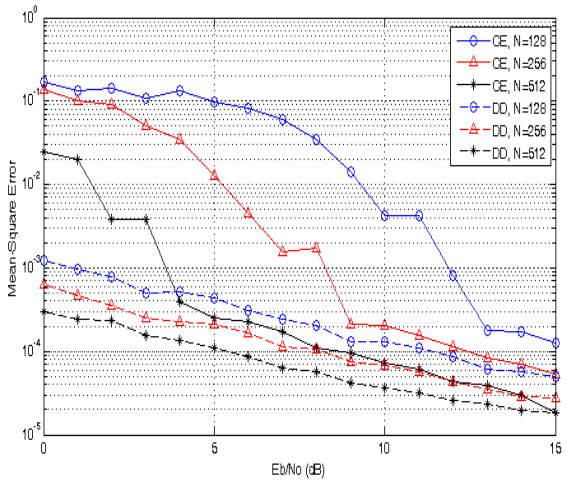


Figure 4. MSE frequency offset estimation performance under flat fading channel conditions: CE – Cyclic Extension; DD – Double-dwell.

The gain achieved by the double-dwell scheme is significantly increased when the channel undergoes frequency-selective fading as shown in Fig. 5. In this case, the channel has five multipath components with equal energy and a maximum delay spread of 75% of the guard time. It is noticed that the cyclic-extension based scheme suffers from the presence of the multipath components while the proposed scheme maintains a good performance.

Finally, we remark that the number of complex multiplications and additions needed by the proposed scheme is $N(1/\epsilon_1 + \epsilon_1/\epsilon_2)$ while a conventional single stage search scheme with the same accuracy as the proposed scheme would require $N(1/\epsilon_2)$ operations. This demonstrates that the proposed scheme reduces the complexity by a factor of $1/(\epsilon_1 + \frac{\epsilon_2}{\epsilon_1})$.

The performance will always be accurate to within $\epsilon_2/2$. Therefore, given a desired maximum error $\epsilon_2/2$, a first stage step size, ϵ_1, that gives the minimum number of search steps, N_s, required for the desired estimation accuracy can be calculated

$$\frac{dN_s}{d\epsilon_1} = \frac{d\left(\frac{1}{\epsilon_1} + \frac{\epsilon_1}{\epsilon_2}\right)}{d\epsilon_1} = 0$$
$$\epsilon_1 = \sqrt{\epsilon_2} \qquad (11)$$

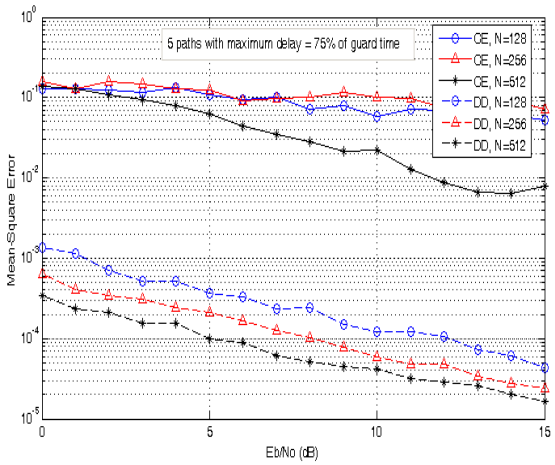


Figure 5. MSE frequency offset estimation performance under frequency-selective fading channel conditions.

V. Conclusion and Future Works

In this paper, a two-stage pilot-based CFO estimation algorithm has been proposed. The algorithm uses a large step size to test for the CFO during the first stage to reduce the complexity. Then, a second search stage is used with a small step size to improve the CFO estimation accuracy. Simulation results demonstrate significant improvement in the MSE of the proposed scheme over both flat and frequency-selective Rayleigh fading channels. Future work will focus on optimizing the number of search stages and step sizes to have further performance improvement and complexity reduction.

References

[1] T. Hwang, C. Yang, G. Wu, S. Li, and G. Li, "OFDM and Its Wireless Applications: A Survey," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1673 – 1694, May 2009.

[2] J. Bingham, "Multicarrier modulation for data transmission: An idea whose time has come," *IEEE Commun. Mag.*, vol. 28, no. 5, pp. 5–14, May 1990.

[3] R. Prasad, *OFDM for Wireless Communication Systems*, Boston: Artech House Inc, 2004.

[4] H. Steendam and M. Moeneclaey, "Analysis and optimization of the performance of OFDM on frequency-selective time-selective fading channels," *IEEE Trans. Commun.*, vol. 47, no. 12, pp. 1811–1819, Dec. 1999.

[5] Y. Mostofi and D. Cox, "Mathematical analysis of the impact of timing synchronization errors on the performance of an OFDM system," *IEEE Trans. Commun.*, vol. 54, no. 2, pp. 226–230, Feb. 2006.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

[6] N. Lashkarian and S. Kiaei, "Class of cyclic-based estimators for frequency-offset estimation of OFDM systems," *IEEE Trans. Commun.*, vol. 48, no. 12, pp. 2139–2149, Dec. 2000.

[7] H. Steendam and M. Moeneclaey, "Analysis and optimization of the performance of OFDM on frequency-selective time-selective fading channels," *IEEE Trans. Commun.*, vol. 47, no. 12, pp. 1811–1819, Dec. 1999.

[8] Y. Li, H. Minn, N. Al-Dhahir, and A. R. Calderbank, "Pilot designs for consistent frequency-offset estimation in OFDM systems," *IEEE Trans. Commun.*, vol. 55, no. 5, pp. 864–877, May 2007.

[9] Y. Li and H. Minn, "Robust and consistent pilot designs for frequency offset estimation in MIMO OFDM systems," *IEEE Trans. Commun.*, vol. 56, no. 10, pp. 1737–1747, October 2008.

[10] J. Zheng and W. Zhu, "An algorithm for calibration of TDS-OFDM carrier frequency offset," *IEEE Trans. Consum. Electron.*, vol. 55, no. 2, pp. 366–370, May 2009.

[11] M. Morelli and U. Mengali, "Carrier-frequency estimation for transmission over selective channels," *IEEE Trans. Commun.*, vol. 48, no. 9, pp. 1580–1589, Sep. 2000.

[12] J. Chen, M. Li, and Y. Kuo, "Adaptive OFDM synchronization algorithm in frequency selective fading channels," *IEEE Trans. Consum. Electron.*, vol. 55, no. 4, pp. 1841 – 1846, Nov. 2009.

[13] W. Chin, "ML estimation of timing and frequency offsets using distinctive correlation characteristics of OFDM signal over dispersive fading channels," *IEEE Trans. Veh. Technol.*, vol 60, no. 2, pp. 444 – 456, Feb. 2011.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Distributed Selection and Optimization of Threshold of Energy Detection for Cooperative Spectrum Sensing

Xiang Lei, Xie Xianzhong, Lei Weijia, Ma Bin
Institute of Personal Communication
Chongqing University of Posts and Telecommunications
Chongqing, China
e-mail: xianglei0727@126.com, xiexianzhong.cn@gmail.com, leiwj@cqupt.edu.cn, mabin@cqupt.edu.cn

***Abstract*—Spectrum detection is the prerequisite for the Cognitive Radio. Energy detection is often used to sense the spectrum hole in Cognitive Radio network, and the threshold plays a vital role. In this paper, a method of setting threshold is proposed. In the low SNR environment, to protect the primary user, each secondary user observes the environment adaptively and then sets the threshold independently. The simulation results prove that, under the method proposed in this paper, the detection performance can be greatly improved, comparing with the traditional one, i.e., the threshold is set by the spectrum broker and all the secondary users have a same fixed threshold.**

Keywords—Cognitive radio; Energy detection; Threshold; Cooperative sensing.

I. INTRODUCTION

In recent years, with the rapid development of radio services, more and more wireless spectrum resources are needed to meet the communication requirements. However, because of the fixed spectrum allocation scheme and the spectrum monopolized principle, the spectrum resources are believed to be scarce resources. Although there are some measures such as Frequency Division Multipexing (FDM), Time Division Multipexing (TDM), Code Division Multiplexing (CDM), etc., to improve the spectrum utilization, but these can not ultimately solve the lack of spectrum resources. As a result, the technology of Cognitive Radio (CR) emerges as the times require. It proposes a opportunists way to share the frequency spectrum, under the premise of not disturbing PU. This method can open the spectrum resources and improve spectrum utilization effectively.

The core ideology of CR is that the unlicensed user (secondary user, SU) senses the radio environment automatically, adjusts the system parameters intelligently to adapt to the changes in the environment, and uses idle spectrum resources to do some communications without disturbing licensed user (primary user, PU). Therefore, idle spectrum sensing is a critical task for CR networks [1][2].

There have been many kinds of classic spectrum sensing technologies, such as energy detection, match filter detection and cyclic-feature detection. However, energy detection is the most commonly used method to estimate if there are any PUs, because of its simple and practical operation.

In the actual communication environments, there are shadowing, fading and other adverse factors which would greatly deteriorate the SU's local decisions. So, the cooperative spectrum sensing is always used in CR network to improve the detection performance, that is, on the base of local detection, all the SUs transmit their local decisions to the spectrum broker through error-free channels, the spectrum broker analyses these data and makes the final decision.

In energy detection, how to set the threshold is very crucial, as the threshold would influence the local decision of SU, and then influence the performance of the system. There have been some studies about how to set the threshold in energy detection [3-7]. In [3], the authors, combining with cooperative sensing, proposed a method to determine the threshold, and the proposed optimal threshold which is decided by spectrum broker minimizes the probability of global error. However, this threshold is not the best one for various SUs who are in different environments. In [4], the authors, combining with the K/N voting rule, discussed the optimal threshold. Depending on K, they only proposed the range of the threshold instead of the exact value. In [5], a tradeoff threshold is proposed, and when the system has a higher requirment of protecting PU, it will set an actual value which is higher than the proposed optimal one to obtain a higher detection probability. In [6], the authors suggested a double-threshold based energy sensing algorithm to improve the performance of the local detection. Under the restriction of the two thresholds, the results of local detection can be more credible, but the authors only considered that all the SUs sensed the idle spectrum over Additive White Gaussian Noise (AWGN) channels, and not taken the fading channels into account. However, in [7], the authors provided a method about how to set a double-threshold in fading channel, and the simulation results proved that the double-threshold proposed can improve the sensing performance.

The detection thresholds mentioned above are all set by the spectrum broker, the spectrum broker bases on some rules to set the threshold uniformly and then distributes it to every SU, each SU, who takes part in cooperation owns a same threshold. However, SUs are in different environments, and they catch PU signals of different intensity. Those ways mentioned above did not take this problem into consideration, but generalized, Thus, all the SUs had a same threshold, which wound reduce the local detection performance.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

In this paper, in order to protect the interests of PU in the low SNR environments, a threshold setting way to improve the performance of local energy detection is proposed. SUs monitor and estimate their radio environments, and then set their own optimal thresholds independently. This optimal one maximizes the difference of SU's the local detection probability and the local false-alarm probability. That is to say, in this way, the system protects the interests of PU to greatest extent, when it allows SUs to share the spectrum resource. The simulation results prove that this method of setting threshold improves the performance of detection significantly, comparing with the traditional method, i.e., the threshold is set by the spectrum broker and all the SUs have a same fixed detection threshold. In addition, the SUs do not need to transmit their signal-to-noise (SNR) radio and some other reliable information to the spectrum broker, so it is also an effective way to save the sensing time and the channel bandwidth.

The remainder of this paper is organized as follows. The energy detection is introduced in Section Ⅱ. The issue about setting threshold uniformly is discussed in Section Ⅲ. The new distributed setting threshold method is described in Section Ⅳ. Simulation results, conclusion and future work are given in Section Ⅴ and Section Ⅵ, respectively.

II. ENERGY DETECTION AND PERFORMANCE CRITERIA

Energy detection is the most commonly used method to estimate if there are any PUs, because of its simple and practical operation.

H_0 denotes PU absent and H_1 denotes PU existing. $y(t)$ is the signal which SU receives. $y(t)$ is denoted by [1]:

$$y(t)=\begin{cases} n(t) & H_0 \\ h(t)\cdot x(t)+n(t) & H_1 \end{cases}, \tag{1}$$

where $x(t)$ is the PU signal, $n(t)$ is the additive white gaussian noise, and $h(t)$ is the complex channel gain of the sensing channel. The SU lets $y(t)$ pass the bandpass filter (BPF) to filter the out-of-band noise and adjacent signals, and then pass the A/D converter, the squarer and the summation device, then the test statistic Y is obtained [2]:

$$Y \to \begin{cases} \chi^2_{2TW} & H_0 \\ \chi^2_{2TW}(2\gamma) & H_1 \end{cases}, \tag{2}$$

where γ is the instantaneous SNR at the secondary node, TW is the product of observation time and interested bandwidth, it is usually written as $m=TW$, and m is an integer. As shown in (2), when PU is absent, Y obeys the central chi-square distribution with $2TW$ degrees of freedom, when the PU is present, Y obeys the noncentral chi-square distribution with $2TW$ degrees of freedom and a non-centrality parameter 2γ.

Letting Y compare to a pre-set threshold λ, SU can decide whether PU is present or not. If Y is bigger than the threshold λ, then SU makes the judgment that PU is working, otherwise, SU believes PU do not occupy this licensed band, and then uses it to do some its own communications. Thus, how to select the threshold is very critical and it influences the local decision of SU immediately, thereby, influences the performance of the system.

In an AWGN environment where the complex channel gain $h(t)$ is constant, the probability of detection, the probability of false-alarm and the probability of missed detection are shown as follows [1]:

$$p_d = p\{Y>\lambda|H_1\} = Q_m\left(\sqrt{2\gamma},\sqrt{\lambda}\right), \tag{3}$$

$$p_f = p\{Y>\lambda|H_0\} = \frac{\Gamma\left(m,\frac{\lambda}{2}\right)}{\Gamma(m)}, \tag{4}$$

$$p_m = p\{Y\le\lambda|H_1\} = 1-p_d, \tag{5}$$

where $\Gamma(.,.)$ is the incomplete gamma function, $Q(\ )$ is the generalized Marcum Q-function, $I_{m-1}(\ )$ is the first modified Bessel function with $m-1$ order. In a fading environment, the complex channel gain $h(t)$ varies with the decline, the SU's average probability of detection is [1]:

$$p_d = \int_x^\infty Q_m\left(\sqrt{2\gamma},\sqrt{\lambda}\right) f_\gamma(x)\,dx, \tag{6}$$

where $f_\gamma(x)$ is the probability distribution function of SNR in the fading environment.

In the actual communication environments, fading and shadowing, etc., would deteriorate the local spectrum sensing performance of the SU, so, multiple SUs are needed to sense the idle spectrum cooperatively, that is the cooperative spectrum sensing. In cooperative spectrum sensing, each SU who takes part in collaboration makes a binary judgment according to the local observation (0 or 1, 0 stands for the absence of PU, and 1 stands for the existence of PU), and then sends the decision to the spectrum broker through ideal channels, the spectrum broker applies the classic K/N voting rule (when K=N, the rule is the AND rule; when K=1, the rule is OR rule) to fuse all the results, and then makes the final decision [2]:

$$\Omega = \sum_{i=1}^{N} D_i \begin{cases} \ge K & H_1 \\ \le K & H_0 \end{cases}, \tag{7}$$

where N is the number of SUs who participate in the cooperation, D_i is the local decision of i th SU. When the number of the SU whose decision is 1, is more than K, the final result is H_1, that is to say the spectrum broker would believe PU is presence, otherwise, the result is H_0, the PU is absence. Under this voting rule, the false-alarm probability of cooperative spectrum sensing and the detection probability of cooperative spectrum sensing is [8]:

$$Q_f = \sum_{j=K}^{N} \sum_{\sum D_i=j} \prod_{i=1}^{N} \left(p_{fi}\right)^{D_i} \left(1-p_{fi}\right)^{1-D_i}, \tag{8}$$

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

$$Q_d = \sum_{j=K}^{N} \sum_{\sum D_i = j} \prod_{i=1}^{N} (p_{di})^{D_i} (1-p_{di})^{1-D_i} , \quad (9)$$

where p_{fi} and p_{di} are the false-alarm probability and the detection probability of i th SU respectively.

III. The Issue about Setting Threshold Uniformly by Spectrum Broker

How to set the threshold in energy detection is very critical. The threshold can influence the false-alarm probability and the detection probability at the same time, when it is set too high, the false-alarm probability would reduce and so does the detection probability. The decrease of the false-alarm probability would increase the radio frequency spectrum utilization, but the decrease of the detection probability would increase the probability of disturbing PU.

As shown in Figure 1, SU1, SU2 and SU3 are all affected by shadowing; SU4 is out of the coverage of the PU transmitter. They can not capture the PU signal no matter it exists or not, and then, they may make error decisions. Although collaborative spectrum sensing can be applied to ameliorate the performance, but the authors of [8] have proved that collaborative spectrum sensing can do little improvement to the performance in the case that secondary nodes are all in harsh environments. In some low SNR environments, SUs would be easy to interfere PU, but the precondition of CR network is that SUs share the spectrum resources without bothering PU. So, In some low SNR environments, especially, in some systems which need to put the interests of PU to the first place, more attention must be paid to protect the interests of PU.

In [3], the authors based on collaborative spectrum sensing, and proposed a method to set the threshold of energy detection. In that way, the threshold (λ^*) minimizes the total error probability (the sum of the missed detection probability and the false-alarm probability). Shown as the following equations:

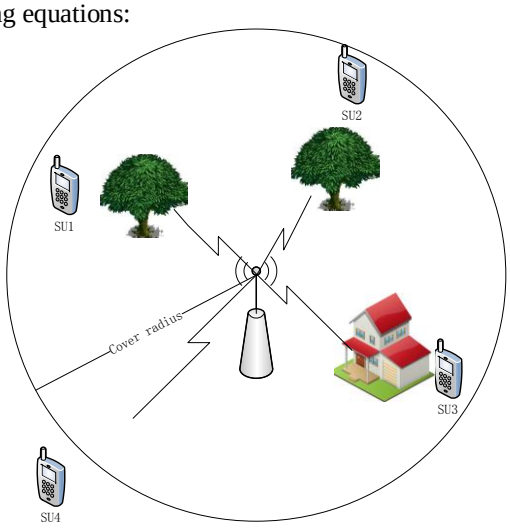


Figure 1. The actual sensing environment

$$\begin{aligned} \lambda^* &= \arg\min_{\lambda}(Q_f + Q_m) \\ &= \arg\min_{\lambda}(Q_f + 1 - Q_d) \\ &= \arg\min_{\lambda}(Q_f - Q_d) \\ &= \arg\max_{\lambda}(Q_d - Q_f) \\ &= \arg\max_{\lambda}(Q_{diff}) \end{aligned} , \quad (10)$$

where $Q_m = 1 - Q_d$ is the system missed detection probability, $Q_{diff} = Q_d - Q_f$ is the difference of the system detection probability and the system false-alarm probability.

Six SUs are assumed to join in the collaborative spectrum sensing, and their SNR are -10dB, -5dB, -3dB, -1dB, 0dB, 1dB respectively, the OR rule is applied to fuse data at the spectrum broker. Letting threshold be the abscissa, the difference of the detection probability and the false-alarm probability (Q_{diff}) be the ordinate, the simulation figure about the relationship between the threshold and the difference of detection probability and false-alarm probability can be obtained, shown as Figure 2. From Figure 2, it is easy to see that the difference of the detection probability and the false-alarm probability varies with the threshold, and a optimal one (λ^*), which maximizes the difference of the detection probability and the false-alarm probability can be obtained. That is to say, according to [3], when the SUs who take part in cooperation all let λ^* be their threshold of energy detection, it would provide the uttermost protection of PU, comparing with some other threshold setting methods.

However, because of the adverse factors such as interference, noise and temperature, the SUs who join in cooperation are in different environments. The threshold mentioned in [3] is set by spectrum broker, and then the spectrum broker distributed it to SUs uniformly, but this threshold is not the optimal one to each SU who is with

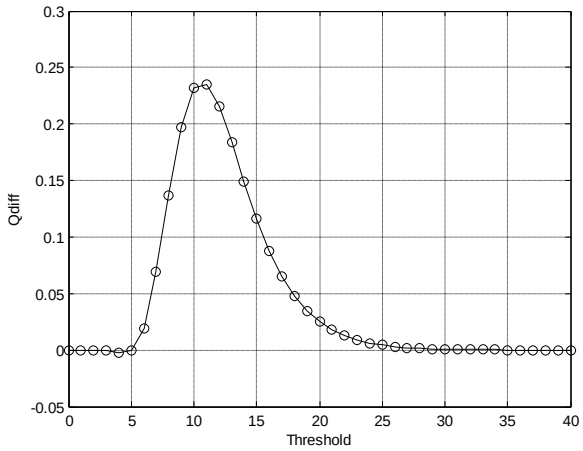


Figure 2. The relationship between the threshold and difference of detection probability and false-alarm probability

different SNR.So, in order to provide better protection to PU, SUs according to their environments, set their own thresholds by maximizing the difference of the detection probability and the false-alarm probability [9], and then make their own optimal decisions independently, that is the distributed setting threshold algorithm. In some harsh environments, this distributed setting threshold method can improve the detection probability, and then achieve the purpose of protecting PU.

IV. THE NEW DISTRIBUTED SETTING THRESHOLD METHOD

We will present the new distributed setting threshold method.

A. The Setting Threshold Method

The following gives a method to set the threshold of energy detection. SUs set their own thresholds according to their environments, and then make their own optimal decisions independently, after that, transmit their decisions to the spectrum broker to complete collaborative spectrum sensing. If the distributed setting threshold method is applied, SUs do not need to convey the SNR and some other reliable informations to the spectrum broker in advance [9][10], and do not need to wait for the uniform fixed threshold which is made by the spectrum broker to finish the energy detection. Obviously, this method can not only save the sensing time, but the bandwidth, more importantly, it can improve the system detection probability, reduce the risk of disturbing PU.

The number of secondary users who participate in collaborative spectrum sensing is N, the SNR of i th SU is assumed to be γ_i $(i=1,2...N)$, the instantaneous detection probability and the instantaneous false-alarm probability of i th SU are shown as follows:

$$p_{di} = Q_m\left(\sqrt{2\gamma_i}, \sqrt{\lambda_i}\right), \tag{11}$$

$$p_{fi} = \frac{\Gamma\left(m, \frac{\lambda_i}{2}\right)}{\Gamma(m)}, \tag{12}$$

where λ_i is the threshold of i th SU. The difference of the detection probability and the false-alarm probability of i th SU is:

$$\begin{aligned} p_{diff} &= p_{di} - p_{fi} \\ &= Q_m\left(\sqrt{2\gamma_i}, \sqrt{\lambda_i}\right) - \frac{\Gamma\left(m, \frac{\lambda_i}{2}\right)}{\Gamma(m)} \end{aligned}. \tag{13}$$

In order to maximize P_{diff}, we take the derivative with respect to λ_i on the both sides of (13), and let it be zero,

$$\begin{aligned} \frac{\partial p_{diff}}{\partial \lambda_i} &= \frac{\partial\left(p_{di} - p_{fi}\right)}{\partial \lambda_i} \\ &= \frac{\partial p_{di}}{\partial \lambda_i} - \frac{\partial p_{fi}}{\partial \lambda_i} \\ &= 0 \end{aligned}, \tag{14}$$

that is:

$$\frac{\partial p_{di}}{\partial \lambda_i} = \frac{\partial p_{fi}}{\partial \lambda_i}, \tag{15}$$

where,

$$\frac{\partial p_{fi}}{\partial \lambda_i} = -\frac{1}{(m-1)!}\frac{\lambda_i^{m-1}}{2^m} e^{-\frac{\lambda_i}{2}}. \tag{16}$$

$$\frac{\partial p_{di}}{\partial \lambda_i} = -\frac{\lambda_i^{\frac{m-1}{2}}}{2\left(2\gamma_i\right)^{\frac{m-1}{2}}} \exp\left(-\frac{\lambda_i + 2\gamma_i}{2}\right) I_{m-1}\left(\sqrt{2\lambda_i\gamma_i}\right). \tag{17}$$

Putting the (16), (17) into (15), and then the optimal threshold which maximize the difference of the detection probability and the false-alarm probability of i th SU can be obtained.

In Rayleigh fading channel, the signal envelope obeys the Rayleigh distribution, the probability density function of γ is:

$$f(\gamma) = \frac{1}{\bar{\gamma}} \exp\left(-\frac{\gamma}{\bar{\gamma}}\right), \gamma \geq 0. \tag{18}$$

Putting (18) into (6), the average detection probability of i th SU can be obtained:

$$\begin{aligned} \overline{p_{di\,Ray}} = {} & e^{-\frac{\lambda_i}{2}} \sum_{k=0}^{m-2} \frac{1}{k!}\left(\frac{\lambda_i}{2}\right)^k + \left(\frac{1+\bar{\gamma}_i}{\bar{\gamma}_i}\right)^{m-1} \\ & \times \left(e^{-\frac{\lambda_i}{2\left(1+\bar{\gamma}_i\right)}} - e^{-\frac{\lambda_i}{2}} \sum_{k=0}^{m-2} \frac{1}{k!}\left(\frac{\lambda_i \bar{\gamma}_i}{2\left(1+\bar{\gamma}_i\right)}\right)^k \right) \end{aligned}. \tag{19}$$

According to (19),

$$\frac{\partial \overline{p_{di\,Ray}}}{\partial \lambda_i} = \frac{1}{2\left(1+\bar{\gamma}\right)} \left(\frac{\Gamma\left(m-1, \frac{\lambda}{2}\right)}{(m-2)!} - \overline{p_{di\,Ray}} \right). \tag{20}$$

Putting (20) into (15), the optimal threshold of i th SU over Rayleigh channel can be obtained.

B. Performance Analyse

SUs set their own thresholds independently according to their environments, and then make their own optimal decisions independently, after that, transmit their decisions to spectrum broker to complete collaborative spectrum detection. Because the issue we study is how to set the threshold to achieve the goal, and the goal is that we not only allow SUs to share spectrum resources, but protect the interests of PU as far as possible. So we apply the OR rule to

fuse data in spectrum broker, and devote to improve the probability of detection.

If the SNR of each SU is assumed to be a instantaneous value, and SUs use the independent setting threshold method to complete energy detection, the system detection probability is:

$$\begin{aligned} Q_d &= 1-\prod_{i=1}^{N}\left(1-p_{di}\right) \\ &= 1-\prod_{i=1}^{N}\left(1-Q_m\left(\sqrt{2\gamma_i},\sqrt{\lambda_i}\right)\right) \end{aligned}, \quad (21)$$

where λ_i and γ_i is the threshold and the instantaneous SNR of i th SU respectively.

If the signal envelope obeys the Rayleigh distribution, and SUs use the independent setting threshold method to complete energy detection, system detection probability is:

$$\begin{aligned} Q_d &= 1-\prod_{i=1}^{N}\left(1-\overline{p}_{di\ Ray}\right) \\ &= 1-\prod_{i=1}^{N}\left(\begin{array}{l} 1-e^{-\frac{\lambda_i}{2}}\sum_{k=0}^{m-2}\frac{1}{k!}\left(\frac{\lambda_i}{2}\right)^k+\left(\frac{1+\overline{\gamma_i}}{\overline{\gamma_i}}\right)^{m-1}\times \\ \left(e^{-\frac{\lambda_i}{2\left(1+\overline{\gamma_i}\right)}}-e^{-\frac{\lambda_i}{2}}\sum_{k=0}^{m-2}\frac{1}{k!}\left(\frac{\lambda_i\overline{\gamma_i}}{2\left(1+\overline{\gamma_i}\right)}\right)^k\right) \end{array}\right) \end{aligned}, \quad (22)$$

where λ_i and $\overline{\gamma_i}$ is the threshold and the average SNR of i th SU respectively.

Over AWGN channel, each SU owns a same SNR value, that is to say $\gamma_1=\gamma_2=.....=\gamma_N=\gamma$, and they can get a same optimal threshold value (λ'), through the independent setting threshold method, and the system detection probability over AWGN channel is:

$$\begin{aligned} Q_d &= 1-\prod_{i=1}^{N}\left(1-p_{di}\right) \\ &= 1-\left(1-Q_m\left(\sqrt{2\gamma},\sqrt{\lambda'}\right)\right)^N \end{aligned}. \quad (23)$$

V. Simulation Results

The number of the SUs who participate in collaborative spectrum sensing is assumed to be $N=6$, and the simulations are did to compare the unified setting threshold method in [3] with the distributed setting threshold method (the algorithm mentioned in this paper) over 10 kinds of environments (the average SNR is -5dB，-4dB，-3dB，-2dB，-1dB，0dB，1dB，2dB，3dB，4dB respectively). The relationships of their system detection probability are shown as Figure 3，Figure 4 and Figure 5.

Figure 3 shows the relationship between the system detection probability and the average SNR under the two methods. From Figure 3, under the distributed setting threshold method, the system detection probability is improved significantly, that is to say, the method proposed in this paper can provide better protection to PU system, comparing with the unified setting threshold method.

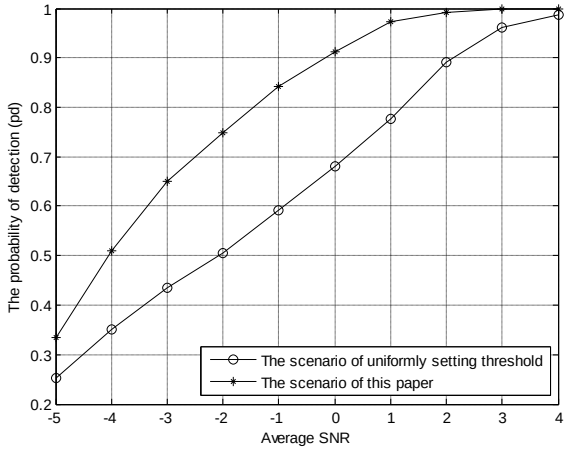


Figure 3. Considering the instantaneous signal-to-noise, the relationship between global detection probability and the average signal-to-noise

Figure 4 and Figure 5 show the simulations over Rayleigh channel and additive white gaussian noise (AWGN) channel respectively. Form the figures, the distributed setting threshold method proposed in this paper is equally applicable to the Rayleigh channel and AWGN channel, this method can also improve the system detection probability, reduce the risk of bothering PU, and achieve the goal of protecting PU.

VI. Conclusion and Future Works

When all the SUs are in low SNR environments, collaborative spectrum sensing can do little improvement to the performance of the system, in this case, we should devote to improve the local detection performance of SU, and then improve the global detection performance. In this paper, each SU according to its environment sets the most suitable threshold independently, and then makes the optimal local decision. The simulation results prove that, under the method proposed in this paper, the local detection

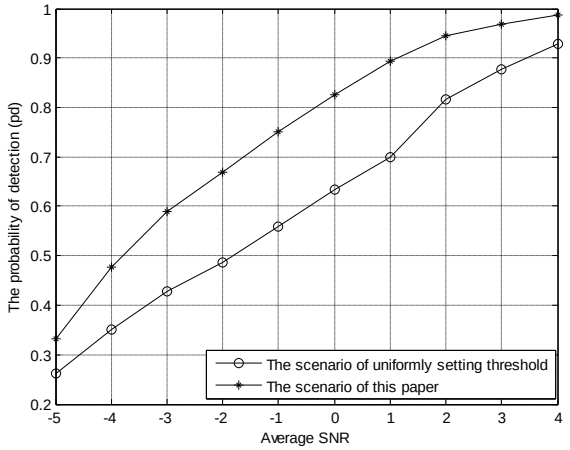


Figure 4. Over Rayleigh channel, the relationship between global detection probability and the average signal-to-noise

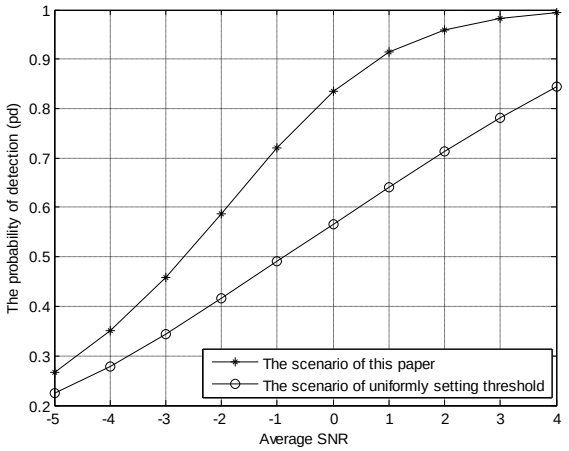


Figure 5. Over AWGN channel, the relationship between global detection probability and the average signal-to-noise

probability is improved, after collaborative detection, the system detection probability is improved too; so, the risk of disturbing PU is reduced. In future work, we will devote to find a more effective way to improve the detection performance of system.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (60872037), the Science & Technology Research Program of the Municipal Education Commission of Chongqing of China (KJ110530) and the Natural Science Foundation of Chongqing (CSTC 2010BB2415, CSTC 2011JJA40006).

REFERENCES

[1] X. Z. Xie, "Cognitive Radio tecnology and application," Beijing: Pulishing House of Electronics Industry, 2008.

[2] T. Yand and H. Arslan, "A survey of spectrum sensing algorithms for Cognitive Radio applications," IEEE Communications Surveys and Tutorials, Vol. 11(1), pp. 116-130, May 2009.

[3] W. Zhang, K. M. Ranjan, and B. L. Khaled, "Optimization of cooperative spectrum sensing with energy detection in Cognitive Radio networks," IEEE Transactions on Wireless Communication, Vol. 8(12), pp. 5761-5766, 2009.

[4] Y. X. Liu, C. Y. Zeng, and H. J. Wang, "Energy detection threshold optimization for cooperative spectrum sensing," 2010 International Conference on Advanced Computer Control (ICACC), pp. 27-29, March 2010.

[5] S. J. Xie, L. F. Shen, and J. S. Liu, "Optimal threshold of energy detection for spectrum sensing in Cognitive Radio," 2009 International Conference on Wireless Communications and Signal Processing (WCSP), pp. 13-15, November 2009.

[6] J. Zhu, Z. G. Xu, and F. R. Wang, "Double threshold energy detection of cooperative spectrum sensing in Cognitive Radio," IEEE CrownCom 2008, pp. 15-17, May 2008.

[7] Alnomay and I.S. Arrabeiah, "Performance analysis of an m,n energy detector in fading environment with double threshold," 2011 13th International Conference on Advanced Communication Technology (ICACT), pp. 13-16, February 2011.

[8] Y. Zheng, X. Z. Xie, and L. L. Yang, "Cooperative spectrum sensing based on SNR comparison in fusion center for Cognitive Radio," 2009 International Conference on Advanced Computer Control, Singapore, pp. 212-216, January 2009.

[9] J. J. Han and J. Li, "Determination of threshold for energy detection in Cognitive Radio sensor networks," China Communications, Vol. 8(1), pp. 14-19, 2011.

[10] Q. Liu, J. Gao, and L. S. Chen, "Optimization of energy eetection based cooperative spectrum sensing in Cognitive Radio networks," 2010 International Conference on Wireless Communications and Signal Processing (WCSP), pp. 21-23, October 2010.

[11] Y. Zheng, X. Z. Xie, and L. L. Yang, "Cooperative spectrum sensing based on SNR comparison in fusion center for Cognitive Radio," 2009 International Conference on Advanced Computer Control, Singapore, pp. 212-216, January 2009.

[12] V. V. Hiep and K. Insoo, "Cooperative spectrum sensing with collaborative users using individual sensing credibility for Cognitive Radio network," IEEE Transactions on Consumer Electronics, Vol.57(2), pp. 320-326, 2011.

[13] Y. H. Zeng, Y. C. Liang, and A. T. Hoang, "A review on spectrum sensing for Cognitive Radio:challenges and solutions," EURASIP Journal on Advances in Signal Processing, pp. 1-15, 2010.

[14] A. Saman, T. Chintha and J. Hai, "Relay based cooperative spectrum sensing in Cognitive Radio network," IEEE , pp. 1-5, doi:10.1109/GLOCOM. 2009. 5425802.

[15] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for Cognitive Radio applications," IEEE Communtions Surveys & Tutorials, Vol. 11(1), pp. 116-130, 2009.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Joint User Scheduling and Link Adaptation for Distributed Antenna Systems in Multi-Cell Environments with Imperfect CSI

Ramiro Sámano-Robles, Eduardo Castañeda Trujillo, and Atílio Gameiro
Instituto de Telecomunicações, Campus Universitário, Aveiro, 3810-193, Portugal
emails:ramiro@av.it.pt, ecastaneda@av.it.pt,amg@ua.pt

***Abstract*—This paper proposes a novel management algorithm for distributed antenna systems (DASs) that exploits the spatial diversity of the distributed architecture in order to schedule (over the same radio resource) as many transmissions as possible with the most appropriate modulation and coding schemes (MCSs). This goal is achieved by implementing a joint user scheduling and link adaptation algorithm (including power control and adaptive modulation and coding) that allows for an appropriate management of intra-cell interference. The algorithm provides the optimum set of scheduled users, the optimum serving nodes, the transmit power levels, and the MCSs that maximize the capacity of the system. In comparison with conventional approaches, where the objective is to maximize the signal-to-interference-plus-noise ratio (SINR) of each user, in this paper the target is to satisfy a given SINR value that ensures the transmission of the chosen MCS with a particular value of block-error-rate (BLER). To achieve this goal, an iterative optimization scheme is proposed in which the set of scheduled users, the power levels, and the MCSs are modified according to channel and interference conditions. A novel method for the calculation of outer-cell interference in multi-cell configurations is also proposed. Imperfect channel state information is used throughout the system-level simulation work. Simulation results show considerable gains in terms of throughput and reduced power consumption per node when compared to conventional systems, thereby making the proposed algorithm suitable for green energy solutions.**

***Index Terms*—Distributed antenna systems, power control, link adaptation, scheduling.**

I. INTRODUCTION

Future wireless networks will make use of advanced algorithms to cope more efficiently with harsh propagation conditions and increasing bandwidth demands. In addition, networks need to be energy efficient and reduce as much as possible dangerous emissions to comply with future regulations regarding health safety and green energy. Over the last few years, multiple antenna technology (also known as multiple-input multiple-output or MIMO) has attracted lots of attention from the research community as a good candidate for boosting the performance of future wireless networks [1]. MIMO systems have the ability to increase the capacity of wireless channels without the need of using additional bandwidth for data transmission [1]. However, due to size and space limitations of user terminals and base stations (BSs) MIMO suffers from the problem of high correlation between the signals of the antenna elements. A solution to this problem can be found in the area of distributed antenna systems (DAS). As compared to co-located antenna systems (CAS), where all antennas are co-located at the central BS, in DASs the antennas or nodes are geographically distributed within the cell [2], thereby reducing access distance to the user and minimizing correlation problems.

Distributed antenna systems were conventionally studied as simple signal relay solutions to improve coverage in indoor locations [3]. However, over the last years, distributed systems have been investigated under more advanced MIMO and multiuser detection schemes. The capacity of DASs with CDMA (code division multiple access) in single cell scenarios has been investigated in [4]. The authors found that capacity gains can be achieved in the down-link by simple selecting for transmission the antenna with the best conditions. By contrast, uplink capacity was maximized by using multiple antenna processing (i.e., multi-user detection). Focusing also on capacity analysis, the work in [2] has proposed a down-link DAS multi-cell scenario with a single user. Two transmission schemes were analyzed: blanket, in which all the antennas assist in the transmission process, and antenna selective, where only the antenna with the lowest path-loss value is selected for transmission. Perfect knowledge of channel state information (CSI) at the transmitter and/or at the receiver was assumed in the analysis. The antenna selective scheme was shown to provide the best performances. Optimum power allocation for DAS in multi-cell environments with a single user has been addressed in [7] and [8].

Despite this extensive work on the physical layer of distributed antenna systems, cross-layer issues such as the design of channel-aware scheduling and resource management algorithms remains relatively unexplored. To partially fill this gap, the work in [5] has analyzed two basic schedulers: round robin and maximum-carrier-to-interference (MCI) for the down-link of distributed systems under different values of traffic load and transmit power. The study has concluded that antenna selective schemes provide considerable gain margins as compared to other solutions, particularly when using round robin scheduling. Relatively less gains were reported in the case of MCI scheduling due to its multiuser diversity properties. Improving on this previous work, a novel packet scheduler for the down-link of DAS using power control has been proposed in [6].

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

This solution aims to select a different user for each distributed antenna and then optimize the antenna power levels in an iterative way in order to comply with a prescribed signal-to-interference-plus-noise ratio (SINR) for each scheduled user. The results showed that the algorithm provides considerable gains in terms of packet throughput that escalate with the number of distributed antennas/nodes inside the central cell.

This paper proposes a further improvement over the algorithm presented in [6] by using different thresholds for different modulation and coding schemes (i.e., adaptive modulation and coding). Therefore, the algorithm also attempts to schedule a different user attached to each one of the nodes in the cell. Each node initially selects the user with the higher channel gain and attempts transmission with the highest possible modulation and coding scheme. Then, an iterative algorithm is used to adapt the transmit power of each node and its associated user in order to satisfy the SINR requirement of the selected modulation and coding scheme. If at the end of this iterative phase the SINR conditions of the scheduled users are not satisfied, then either the modulation and coding schemes or the set of scheduled users are modified. This scheme is repeated until the conditions of all the scheduled users in the cell are satisfied. In this way, the set of optimum scheduled users, their transmit power levels, and modulation and coding schemes that maximize system capacity are obtained for a particular time-slot of the system. In order to calculate outer-cell interference, the results of the power levels calculated in previous simulation runs are reused in the outer-cells to replicate in a better way the behavior of the algorithm at the system level. The results show that the proposed algorithm improves considerably the throughput of the system using lower values of transmit power per node, thereby being suitable for green energy solutions in future deployments.

The structure of this paper is as follows. Section II describes the multi-cell deployment scenario and the propagation and signal models to be used. Section III describes the proposed algorithm and the optimization techniques. Section IV presents the results of the simulation work. Finally, Section V draws the main conclusions of the paper.

II. System model

Consider the hexagonal multi-cell distributed antenna system depicted in Fig. 1 with $I+1$ cells: one central cell ($i = 0$) which will be the main target of analysis, and I surrounding cells ($i = 1, \ldots, I$), which will be used as simple sources of outer-cell interference. Only one tier of surrounding cells will be used (i.e., $I = 6$). Each hexagonal cell has a radius R and consists of a total of $L + 1$ radiation nodes: one located at the center of the cell ($l = 0$), and L distributed nodes ($l = 1, \ldots, L$) located at a distance D_r from the center of the cell. The distributed nodes are spaced at uniform angles given by $\theta_l = \frac{2(l-1)\pi}{L}$. A conventional cellular system with one centralized node can be characterized by substituting $L = 0$ in all the expressions in this paper. It is also assumed that the distributed nodes are connected, via a dedicated link such as a coaxial cable or optical fibre, to the node at the center of the cell where all decisions for user scheduling and power allocation are taken.

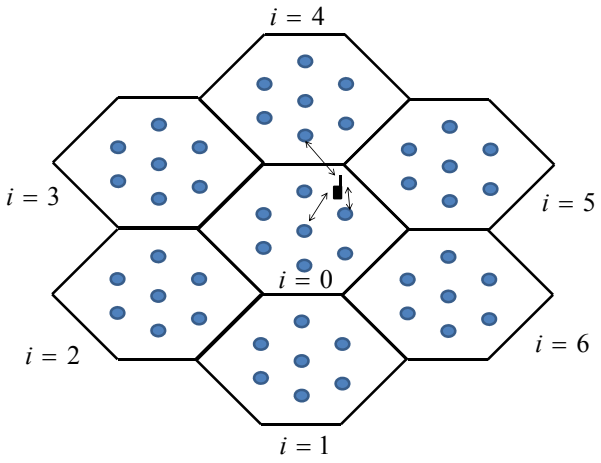


Fig. 1. Cellular Architecture for evaluation of DAS.

All the transmissions in the network are organized in time-slots. Each transmission can use one of the M possible modulation and coding schemes (MCSs). A set of J potential users is considered to be randomly deployed in the central cell of analysis every time slot of the system. The channel between user j and the l-th node of the i-th cell of the network will be denoted by $h_{l,i,j}$. Channel envelopes of different users and different distributed nodes are assumed to be statistically independent and with Rice distribution described by the parameter K. This means that $h_{l,i,j}$ will be modeled as a complex Gaussian variable with mean μ and variance σ^2, i.e. $h_{l,i,j} \sim \mathcal{CN}(\mu, \sigma^2)$, where $K = \frac{\mu^2}{\sigma^2}$. The channels are affected by a propagation path-loss model defined by [9]:

$$L_{dB}(l,i,j) = 20\log_{10}(d_{l,i,j}) + 44.3 + 20\log_{10}\left(\frac{f}{5.0}\right), \quad (1)$$

where $d_{l,i,j}$ is the distance (in meters) between user j and the l-th node of the i-th cell of the network, and f is the operational frequency in GHz. Shadowing is also considered using a log-normal distribution with parameter $\sigma_s = 3dB$. The signal transmitted by the l-th node of the i-th cell will be denoted by $\mathbf{s}_{i,l} = [s_{i,l}(0), \ldots, s_{i,l}(S-1)]^T$, where S is the number of symbols and $(\cdot)^T$ is the vector transpose operator. Assuming that the transmitted symbols have unitary power (i.e., $E[\mathbf{s}_{i,l}^H \mathbf{s}_{i,l}] = 1$, where $E[\cdot]$ is the expectation operator and $(\cdot)^H$ is the hermitian transpose operator) and the transmit power of the l-th node in cell i is given by $P_{l,i}$, then the signal received by user j can be written as:

$$\mathbf{r}_j = \sum_{i=0}^{I} \sum_{l=0}^{L} \sqrt{P_{l,i}} h_{l,i,j} \mathbf{s}_{i,l} + \mathbf{v}_j, \quad (2)$$

where $\mathbf{v}_j = [v_j(0), \ldots, v_j(S-1)]^T$ is the additive gaussian noise with zero mean and unitary variance $v_j(q) \sim$

$\mathcal{CN}(0,\sigma_v^2), q \in \{1,\dots S-1\}$ where $\sigma_v^2 = 1$. The signal-to-interference-plus-noise ratio (SINR) experienced by user j in cell i given the transmission of the l-th node, which is also in the i-th cell, is denoted by $\gamma_{l,i,j}$ and can be mathematically written as:

$$\gamma_{l,i,j} = \frac{P_{l,i}|h_{l,i,j}|^2}{1+\sum_{n=0;n\neq l}^{L} P_{n,i}|h_{n,i,j}|^2 + \upsilon_{i,j}}, \quad j \in \mathcal{U}_i \qquad (3)$$

where $\upsilon_{i,j} = \sum_{k=0;k\neq i}^{I}\sum_{n=0}^{L} P_{n,k}|h_{n,k,j}|^2$ is the outer-cell interference when user j is incell i, and $\mathcal{U}_i$ is the set of users located in the coverage area of cell i. Since all the decisions for resource allocation, user scheduling and power control will be taken at the central node, the available channel state information is potentially inaccurate. In this paper we will assume that the central node has perfect knowledge of long term channel statistics, such as average power and the line-of-sight component of the Rician-distributed channels, and imperfect knowledge of the random fading component. The channel variable available at the central node will be denoted by $\hat{h}_{l,i,j}$, and the accuracy of the channel state information (CSI) will be characterized by a correlation coefficient defined as $\rho = \frac{E[(\hat{h}_{l,i,j}-\mu)(h_{l,i,j}-\mu)]}{\sigma^2}$. The SINR measured by the central node in the cell will be then given by:

$$\hat{\gamma}_{l,i,j} = \frac{P_{l,i}|\hat{h}_{l,i,j}|^2}{1+\sum_{n=0;n\neq l}^{M} P_{n,i}|\hat{h}_{n,i,j}|^2 + \hat{\upsilon}_{i,j}}, \qquad (4)$$

where $\hat{\upsilon}_{i,j}$ is the estimated outer-cell interference for user j in cell i.

III. Algorithm Description

The main objective of the algorithm proposed in this paper is to multiplex/schedule as many users as possible over the same frequency band while maximizing capacity in the cell. Each user will be attached to each one of the distributed nodes inside the cell (only one user per node). The algorithm aims to optimize the power levels of the nodes as well as their modulation and coding schemes in order to reduce interference and maximize the throughput in the cell. The steps of the algorithm can be described as follows:

1) Simulation is initialized
2) Users are placed in random positions across the central cell.
3) For each one of the distributed nodes in the central cell the best user is selected based on the measured channel gain:

$$u_l = \arg\max_j |\hat{h}_{l,0,j}|, \quad u_l \neq u_n, \quad ,(n,l) \in \{0,\dots,L\}$$

4) For all the selected users the maximum possible modulation and coding scheme is initially selected as well as the maximum transmit power level per node P_{max}.
5) Power levels required to satisfy the SINR of the modulation and coding scheme of each scheduled user (denoted by $\gamma_{l,0,u_l}^{(mcs)}$) are updated using eq.(4):

$$\tilde{P}_{l,0} = \frac{\gamma_{l,0,u_l}^{(mcs)}(1+\sum_{n=1;n\neq l}^{M} P_{n,0}|\hat{h}_{n,0,u_l}|^2 + \hat{\upsilon}_{0,u_l})}{|\hat{h}_{l,0,u_l}|^2},$$

$$P_{l,0} = \min(P_{max}, \tilde{P}_{l_0}), \quad l \in \{0,\dots,L\}$$

6) The actual SINR achieved by each user is obtained based on the updated power levels using eq.(4). If all the users have satisfied their required SINR level then the algorithm jumps to the next step. Otherwise, the user with the highest transmit power requirement must be allocated with a modulation and coding scheme with less SINR requirements. In case there is no other modulation and coding scheme with less SINR requirement then the user and the serving node are dropped from the set of scheduled users/nodes ($P_{l,0} = 0$). The algorithm then goes back to step 4.
7) The power levels of the outer-cells ($P_{l,i}$, $i \neq 0$) are updated using the results of the central cell, and another iteration is started by going back to step 2.

A flowchart of the proposed algorithm describing these steps is shown in Fig. 2.

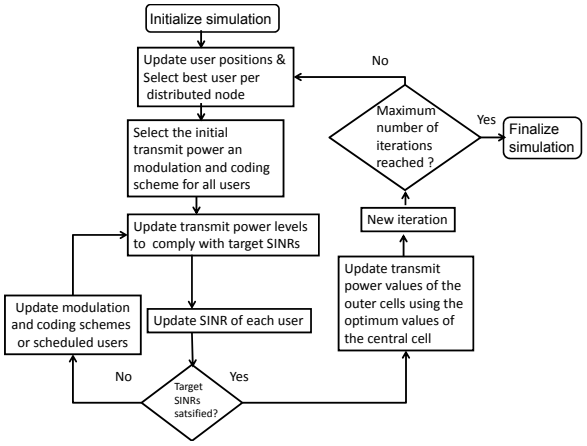


Fig. 2. Flowchart describing the proposed radio resource management algorithm for DASs.

IV. Results

This section presents simulation results that show the benefits of the proposed algorithm. The main metric to be used is throughput (T), which can be defined as the ratio of the total amount of bits successfully transmitted to the total time used in the transmission of that information. In all the simulations, the throughput is calculated by means of look up tables. Once the instantaneous SINR of each user has been calculated, its value is mapped into a look up table with thresholds for each modulation and coding scheme. If the SINR surpasses the threshold of the targeted modulation and coding scheme then the information can be considered as correctly transmitted with a given block error rate (BLER). The modulation and coding schemes and their corresponding thresholds and BLER values are shown in Table I. These modulation and coding schemes correspond to the WiMAX system as given in [10] with a block size of $Q = 7200$ symbols and a frame length of $Fl = 5ms$. The mathematical expression for the instantaneous throughput

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

given that the SINR surpasses the targeted threshold is given by:

$$T = \frac{(1 - BLER)B \cdot Q \cdot R_c}{R \cdot Fl},$$

where R_c is the rate of the turbo code scheme, B is the number of bits per constellation, and $R = 6$ is the repetition coding rate [10]. The results obtained after 10,000 Monte-carlo simulation runs are displayed in Fig. 3 for the average throughput versus various values of the transmit power-to-noise ratio ($\frac{P_{max}}{\sigma_v^2}$) in dBs, using a Rice factor of $K = 10$dB for the particular case of $\rho = 1$, i.e. a system with perfect channel state information. In all simulations, $J = 10$ user terminals, a cell radius of $R = 500m$ and a node position of $D_r = 2/3R$ have been used. It can be observed in Fig. 3 that the proposed algorithm for DAS ($L = 6$) provides a considerable gain in throughput over a conventional cellular system ($L = 0$). To further illustrate these gains, Fig 4 shows the throughput gain of the proposed algorithm with respect to a conventional cellular system. It can be observed that at low values of transmit power the gain reaches almost the maximum theoretical value of $L + 1 = 7$ for MIMO systems, but it reduces the performance for higher values of transmit power to almost 2.5. In terms of power consumption, Fig. 5 shows that the average transmit power per node in DAS using the proposed algorithm is considerably lower (by almost 10dB) than the power consumed by a conventional cellular system without power control. To illustrate the statistical performance of the iterative scheme, Fig. 6 shows the average number of iterations required by the proposed scheme to reach the optimum solution. It can be observed that higher number of iterations are required for low values of transmit power (almost 70), while at low values of power, the number reduces to nearly 50. This means that the proposed algorithm can converge more or less quickly to the desired solution. A method to speed up the performance is by improving the initial conditions of the iterative scheme by figuring out which users will be likely to be dropped, or by making a better guess of the MCS to be used by a given terminal. Since the proposed algorithm also aims to allow simultaneous user transmissions within the cell, Fig. 7 displays the average number of scheduled users per time-slot or TTI (time-transmission-interval). It can be observed that at higher values of power more users can be simultaneously served by the system, reaching a maximum close to 4 scheduled users, which indicates that nearly 50% of nodes are deactivated each TTI. To illustrate the advantages of the algorithm in terms of the usage of higher order MCSs, Fig. 8 shows the average usage of the top three MCSs in Table I, where it can be observed that the algorithm allows more frequent use of these MCSs.

To illustrate the effects of imperfect channel state information, Fig. 9 shows the average throughput performance of the proposed algorithm in DAS and for a conventional cellular system versus different values of the correlation factor ρ using a Rice factor of $K = -\infty$ dB (Rayleigh fading) and a fixed value of transmit power-to-noise ratio of $\frac{P_{max}}{\sigma_v^2} = 95dB$. It can be observed that both schemes can be considerably affected by the effects of imperfect CSI, particularly when the correlation factor is below 0.9. The proposed algorithm for DAS results more affected than a conventional cellular system. At very low values of correlation factor ($\rho < 0.7$), the performance can be slightly worse than the performance of conventional cellular systems with perfect channel state information. Therefore, it is important for the correct operation of the proposed algorithm to have a reliable channel state information to reduce throughput losses in Rayleigh fading channels. Fig. 10 shows the average throughput performance of the proposed algorithm in DAS and for a conventional system versus different values of the correlation factor ρ using a Rice factor of $K = 10$dB. As shown in Fig. 10, both schemes can be affected by the effects of imperfect CSI when the correlation factor is below 0.9. It can be observed that the proposed algorithm is not affected as much as in the case of Rayleigh fading. In fact, the performance is always higher than that of the conventional cellular system. Theses results show that the the proposed algorithm is more robust to the effects of imperfect channel state information in environments with good line-of-sight.

TABLE I
WiMAX MODULATION AND CODING SCHEMES [10].

QPSK 1/3		**QPSK 1/2**		**QPSK 2/3**	
SINR	**BLER**	**SINR**	**BLER**	**SINR**	**BLER**
-1.94	1.00e+0	0.62	1.00e+0	2.67	1.00e+0
-1.74	9.95e-1	0.82	9.45e-1	2.87	9.90e-1
-1.54	8.03e-1	1.02	3.95e-1	3.07	6.76e-1
-1.34	1.79e-1	1.22	2.76e-2	3.27	9.97e-2
-1.14	4.10e-3	1.32	4.13e-3	3.47	6.50e-3
QPSK 3/4		**QPSK 4/5**		**16 QAM 1/3**	
SINR	**BLER**	**SINR**	**BLER**	**SINR**	**BLER**
3.98	1.00e+0	4.66	1.00e+0	3.06	1.00e+0
4.18	9.40e-1	4.86	9.94e-1	3.26	9.14e-1
4.38	3.93e-1	5.06	7.28e-1	3.46	2.58e-1
4.58	3.97e-2	5.26	1.38e-1	3.56	5.72e-2
4.78	3.30e-3	5.46	4.97e-3	3.66	7.15e-3
16 QAM 1/2				**16 QAM 2/3**	
SINR	**BLER**			**SINR**	**BLER**
5.82	1.00e+0			8.47	1.00e+0
6.02	9.94e-1			8.67	9.92e-1
6.22	5.89e-1			8.87	6.67e-1
6.42	4.49e-1			9.07	1.08e-1
6.52	5.70e-3			9.37	3.80e-3
16 QAM 3/4				**16 QAM 4/5**	
SINR	**BLER**			**SINR**	**BLER**
10.18	1.00e+0			11.07	1.00e+0
10.38	8.95e-1			11.27	9.51e-1
10.58	2.79e-1			11.47	3.60e-1
10.78	2.00e-2			11.67	2.42e-2
10.98	1.57e-3			11.77	3.30e-3

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

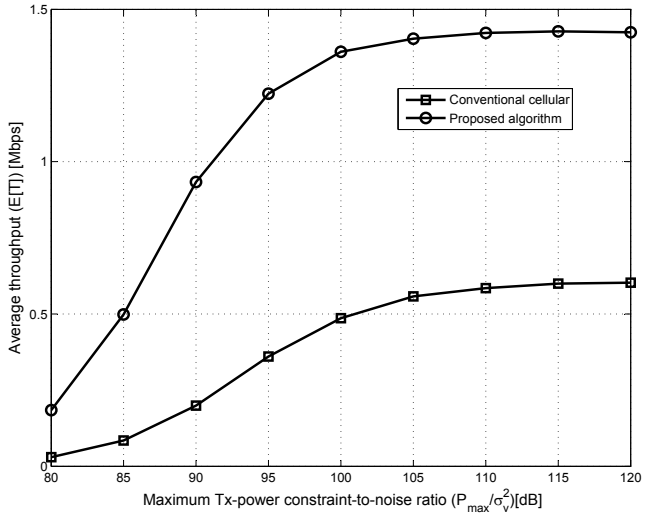


Fig. 3. Average throughput ($E[T]$) vs. maximum transmit power-to-noise ratio $\frac{P_{max}}{\sigma_v^2}$ [dB] for the proposed algorithm in DAS and for conventional cellular systems.

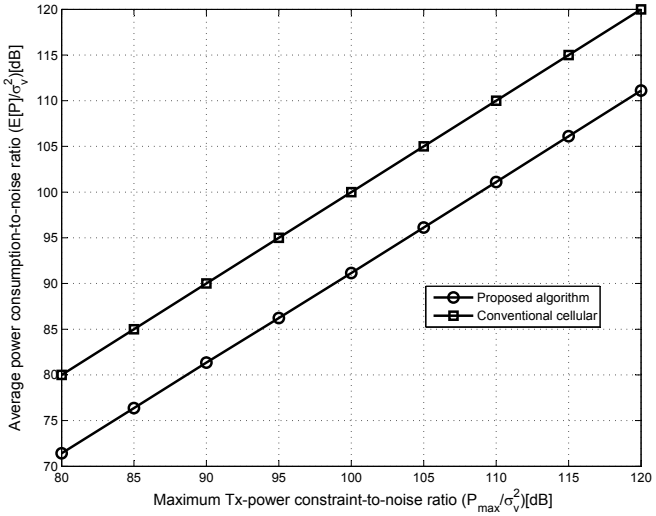


Fig. 5. Average transmit power consumption per node (E[P]) vs. maximum transmit power-to-noise ratio $\frac{P_{max}}{\sigma_v^2}$ [dB] for the proposed algorithm in DAS and for conventional cellular systems.

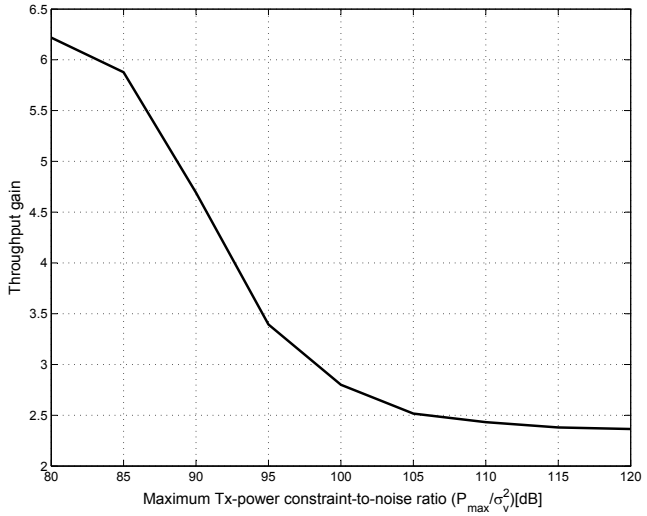


Fig. 4. Average throughput gain vs. maximum transmit power-to-noise ratio $\frac{P_{max}}{\sigma_v^2}$ [dB] for the proposed algorithm in DAS with respect to conventional cellular systems.

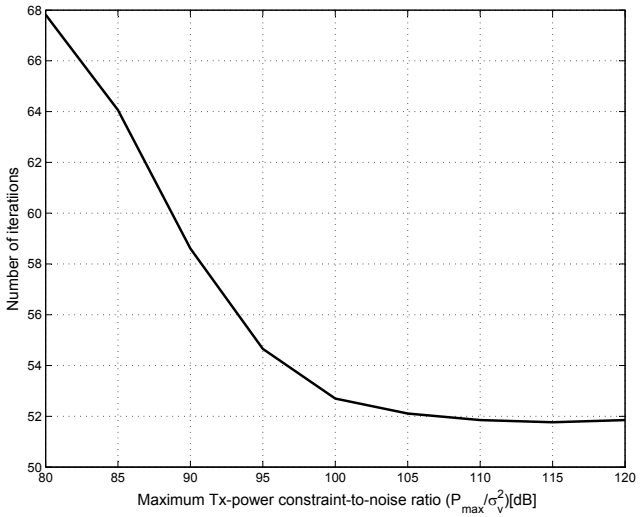


Fig. 6. Average number of iterations vs. maximum transmit power-to-noise ratio $\frac{P_{max}}{\sigma_v^2}$ [dB] for the proposed algorithm in DAS.

V. Conclusions and Future Works

This paper has presented a new algorithm for the optimization of distributed antenna systems that allows the simultaneous transmission of several users attached to different nodes in the cell with controlled power to reduce inter-cell interference and with adaptive modulation and coding. The algorithm shows that by controlling inter-cell interference based on a cross-layer approach, high capacity gains can be achieved by exploiting the spatial diversity of the distributed nodes in the system. In addition, considerable power transmission reduction can be achieved while preserving high throughput gains, particularly in environments with good line-of-sight. This feature makes the algorithm suitable for green energy solutions. The results also show that the maximum gain is close to the theoretical boundary of MIMO systems, which is equal to the number of antennas in the system. It was also observed during the simulation work that users that previously were discriminated due to their poor channel conditions have now more chance to get access to network resources. Analysis of fairness for the proposed algorithm is an interesting future research topic. Future works include the use of beam-forming across different distributed nodes, the extension of the algorithm to the uplink case, and also considering that users have a finite buffer with data to be transmitted.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

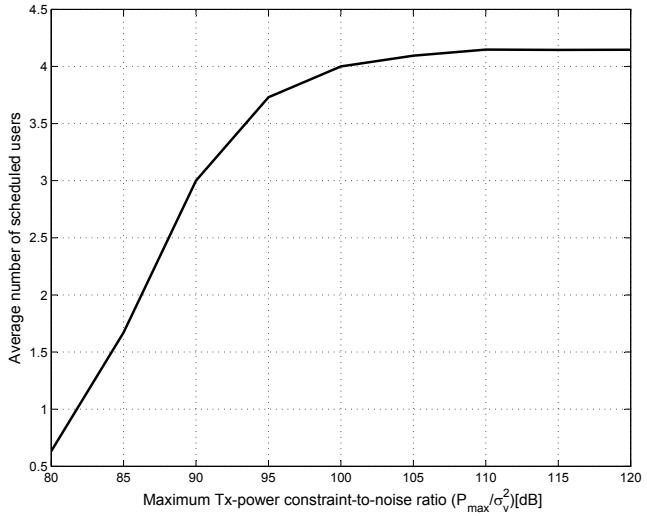


Fig. 7. Average number of scheduled users vs. maximum transmit power-to-noise ratio ($\frac{P_{max}}{\sigma_v^2}$) [dB] for the proposed algorithm in DAS.

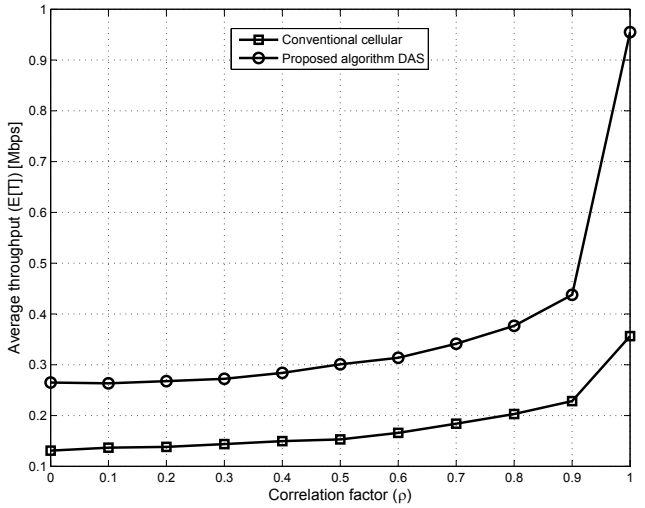


Fig. 9. Average throughput ($E[T]$) vs. correlation factor (ρ) for the proposed algorithm in DAS and for conventional cellular systems using $K = -\infty$dB.

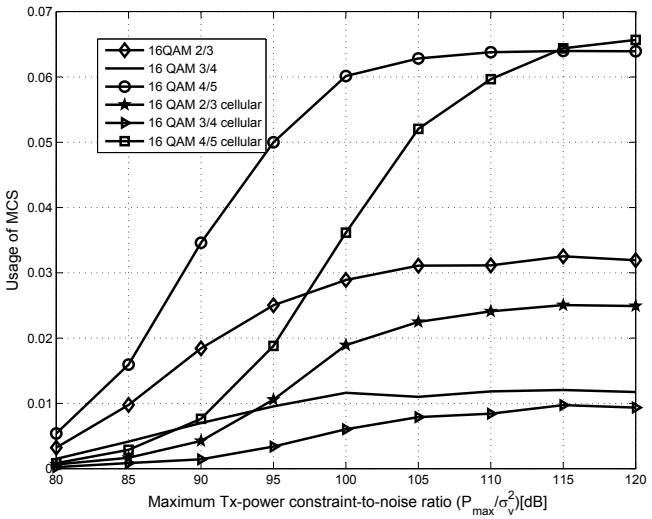


Fig. 8. Average usage of the top three MCSs from table I vs. maximum transmit power-to-noise ratio ($\frac{P_{max}}{\sigma_v^2}$) [dB] for the proposed algorithm in DAS and for conventional cellular systems.

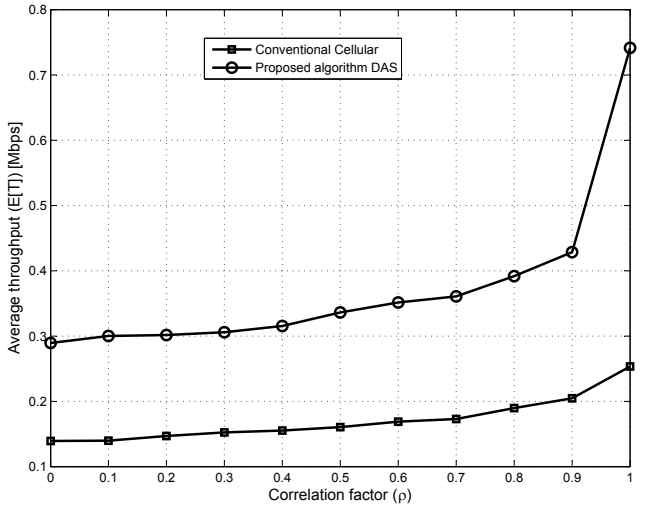


Fig. 10. Average throughput ($E[T]$) vs. correlation factor (ρ) for the proposed algorithm in DAS and for conventional cellular systems using $K = 10$dB.

REFERENCES

[1] A. Goldsmith, S.A. Jaffar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE Journal on Selected Areas in Communications,* vol. 25, no. 5, pp. 684-702, June 2003.

[2] W. Choi, and J.G. Andrews, "Downlink performance and capacity of distributed antenna systems in a multi-cell environment," *IEEE Transactions on Wireless Communications,* vol. 6, no. 1, pp. 69-73, January 2007.

[3] A.M Saleh, A. Rustako, and R. Roman, "Distributed antennas for indoor radio communications," *IEEE Transactions on Communications,* vol. 35, no. 12, pp. 1245-1251, December 1987.

[4] L. Dai, S. Zhou, and Y. Yao, "Capacity analysis in CDMA distributed antenna systems," *IEEE Trans. letters on Wireless Commun.,* vol. 4, no. 6, pp. 2613-2620, November 2005.

[5] R. Samano-Robles and A. Gameiro, "A cross-layer approach to the downlink performance analysis and optimization of distributed antenna systems in multi-cell environments," *1st Int. Conf. on wireless commun., vehicular tech., info. theory, and aerospace&elec. systems tech., 2009, Wireless Vitae,* Aalborg, Denmark, pp. 166-170.

[6] R. Samano-Robles and A. Gameiro "An SINR-based Packet Scheduling Algorithm with Antenna Diversity Selection for Distributed Broadband Wireless Systems," Inivted paper to *Conference on Wireless Personal Mobile Communications,* Japan, March 2009.

[7] W. Feng, X. Zhang, S. Zhou, J. Wang, and M. Xia, "Downlink power allocation for distributed antenna systems in a multi-cell environment," *Int. conf. on wireless commun., net. and mobile comput., wicom* 2009, pp. 1-4.

[8] W. Feng, X. Zhang, S. Zhou, J. Wang, and Minghua Xia, "Downlink power allocation for distributed antenna systems with random antenna layout," *Vehicular technology conference fall* 2009, pp. 1-5.

[9] WINNER deliverable D1.1.2 . Final link level and system-level channel models.

[10] *WiMAX system level evaluation methodology,*WiMAX Forum, V.0.0.1, 2006.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Open API for M2M Applications: What is Next?

Current state and development proposals

Manfred Schneps-Schneppe
Ventspils University College
Ventspils International Radio Astronomy Centre
Ventspils, Latvia
manfreds.sneps@gmail.com

Dmitry Namiot
Lomonosov Moscow State University
Faculty of Computational Mathematics and Cybernetics
Moscow, Russia
dnamiot@gmail.com

***Abstract*—This paper relates to telecommunication standards and describes the current status of open Application Programming Interface for M2M applications as well as proposes some changes and extensions. The European Telecommunications Standards Institute is going to provide open standards for the rapidly growing M2M market. An open specification, presented as an Open API, provides applications with a rich framework of core network capabilities upon which to build services while encapsulating the underlying communication protocols. Services may be replicated and ported between different execution environments and hardware platforms. We would like to discuss the possible extensions for ETSI proposals and describe the add-ons that, by our opinion, let keep telecom development inline with the modern approaches in the web development domain.**

Keywords-m2m; open api; rest; web intents.

I. INTRODUCTION

Machine-to-Machine (M2M) refers to technologies that allow both wireless and wired systems to communicate with other devices of the same ability. M2M uses a device (such as a sensor or meter) to capture an event (such as temperature, inventory level, etc.), which is relayed through a network (wireless, wired or hybrid) to an application (software program), translates the captured event into meaningful information [1].

Considering M2M communications as a central point of Future Internet, European commission creates standardization mandate M/441 [2]. The general objective of the mandate is to ensure European standards that will enable interoperability of utility meters (water, gas, electricity, heat), which can then improve the means by which customers' awareness of actual consumption can be raised in order to allow timely adaptation to their demands.

Our goal is here to propose some new additions for M2M communications, namely, web intents, as add-on for the more traditional REST approach to simplify the development phases for M2M applications. The key advantages are JSON versus XML, asynchronous communications and integrated calls.

Right now, market players are offering own standards for M2M architecture [17].

Figure 1 illustrates the basics of M2M infrastructure (as per Cisco) [3].

The M2M infrastructure includes three primary domains: cloud, network, and edge devices. Each of these domains contains a specific anchor point which conducts the M2M signaling across the infrastructure. The M2M traffic has its own specific characteristics, such as low mobility and offline and online data transmission, which create new challenges for dimensioning the network. Service providers that are trying to customize their networks face the additional challenge of supporting traffic generated from residential and enterprise customer premises equipment (CPE).

Of course, there are several attempts to provide the standard set of software tools for M2M applications. These attempts are well explainable. Because M2M applications are directly linked to hardware devices than the portability of applications, the ability to bring new devices in system etc. become the key factors.

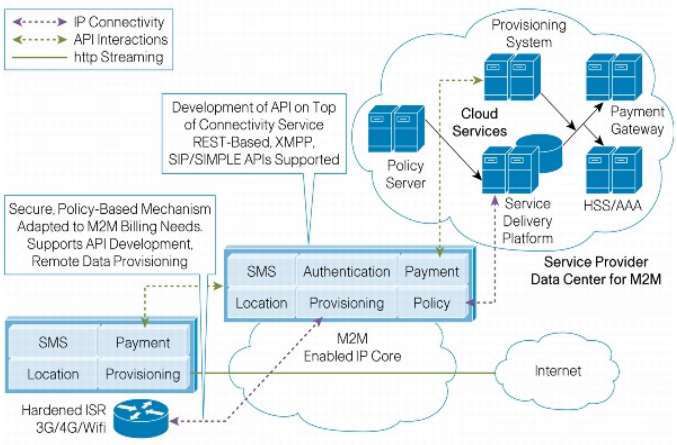


Figure 1. M2M infrastructure (as per Cisco)

Current customized M2M solutions and platforms tend to assume direct connectivity between the M2M core and devices, with no aggregators. However, linking residential and enterprise M2M gateways to an M2M-ready core opens new business models for service providers. M2M gateways can be bypassed when necessary.

In other words, what we can see now it is a growing interest to the M2M middleware.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Also, M2M middleware helps us with heterogeneity of M2M applications. Heterogeneity of service protocols inhibits the interoperation among smart objects using different service protocols and/or API's. We assume that service protocols and API's are known in advance. This assumption prevents existing works from being applied to situations where a user wants to spontaneously configure her smart objects to interoperate with smart objects found nearby [4].

Alcatel [5], for example, proposes the following conceptual view of M2M server (Fig. 2).

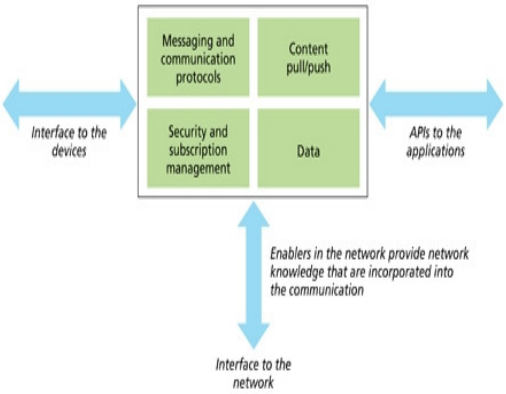


Figure 2. M2M server architecture

The gateway element should be located on the boundary between a wireless network and the Internet network used by application servers to communicate to a device. So, the M2M server can maintain sessions to application servers on one side, and to devices on the other side. In other words, it acts as a bridge, passing information from the application server to appropriate devices.

Web based architecture (or similar to web based) is about the common trend as we see. Many systems are offering for M2M developers the tools that developers are familiar with (e.g., from the previous projects, from the enterprise development, etc.), but the common denominators here are the standard protocols (REST, SOAP over HTTP) and nothing more. In other words, we can see the REST support in many (almost all) M2M frameworks, but the semantic for calls could be (almost always) different.

Of course, ETSI [2] is not the only source for the standardization in M2M area. The 3rd Generation Partnership Project maintains and develops technical specifications and reports for mobile communication systems [15]. The International Telecommunication Union as a specialized agency of the United Nations is responsible for IT and communication technologies. The Telecommunications Standardization Sector (ITU-T), covers the issue of M2M communication via the special Ubiquitous Sensor Networks-related groups [16]. ITU address the area of networked intelligent sensors.

Also, we can see a growing interest to the cloud based M2M systems. For example, Axeda [14] offers cloud for M2M devices, including many traditional elements from the enterprise development world like business rules in orchestrations. [Fig. 3]

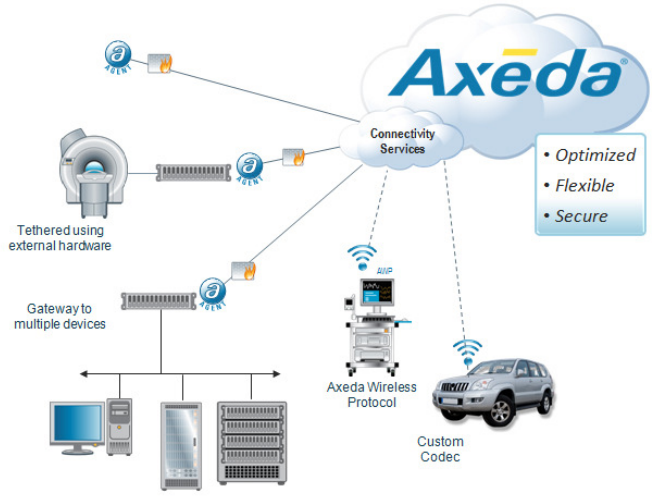


Figure 3. Axeda platform

Note that this system is actually very far from the European standards, despite the fact that it is also based on REST and SOAP as the ETSI standards. But in the same time AT&T has selected it as default M2M platform.

The rest of the paper is organized as follows. Section II contains an analysis of Open API for M2M, submitted to ETSI. In Section III, we offer the never web tool, Web Intents for enhancement of M2M middleware. Sections IV and V are devoted to discussion.

II. OPEN API FROM ETSI

This section describes an Open API for M2M, submitted to ETSI. It is probably the most valuable achievement at this moment.

The OpenAPI for M2M applications developed jointly in Eurescom study P1957 [6] and the EU FP7 SENSEI project [7] makes. The OpenAPI has been submitted as a contribution to ETSI TC M2M [8] for standardization.

Actually, in this Open API, we can see the big influence of Parlay specification. Parlay Group leads the standard, so called Parlay/OSA API, to open up the networks by defining, establishing, and supporting a common industry-standard APIs. Parlay Group also specifies the Parlay Web services API, also known as Parlay X API, which is much simpler than Parlay/OSA API to enable IT developers to use it without network expertise [9].

The goals are obvious, and they are probably the same as for any unified API. One of the main challenges in order to support easy development of M2M services and applications will be to make M2M network protocols "transparent" to applications. Providing standard interfaces to service and application providers in a network independent way will allow service portability [10].

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

At the same time, an application could provide services via different M2M networks using different technologies as long as the same API is supported and used. This way an API shields applications from the underlying technologies, and reduces efforts involved in service development. Services may be replicated and ported between different execution environments and hardware platforms [11].

This approach also lets services and technology platforms to evolve independently. A standard open M2M API with network support will ensure service interoperability and allow ubiquitous end-to-end service provisioning.

The Open API provides service capabilities that are to be shared by different applications. Service Capabilities may be M2M specific or generic, i.e., providing support to more than one M2M application.

Key points for Open API:

- it supports interoperability across heterogeneous transports

- ETSI describes high-level flow and does not dictate implementation technology

- it is message-based solution

- it combines P2P with client-server model

- and it supports routing via intermediaries

At this moment all point are probably not discussable except the message-based decision. Nowadays, publish-subscribe method is definitely not among the favorites approaches in the web development, especially for heavy-loading projects.

Let us name the main Open API categories and make some remarks.

ETSI Open API categories	API contents	Comments
Grouping	A group here is defined as a common set of attributes (data elements) shared between member elements. On practice it is about the definition of addressable and exchangeable data sets.	Just note, as it is important for our future suggestions, there are no persistence mechanisms for groups.
Transactions	Service capability features and their service primitives optionally include a transaction ID in order to allow relevant service capabilities to be part of a transaction. Just for the deploying transactions and presenting some sequences of operations as atomic.	In the terms of transactions management Open API presents the classical 2-phase commit model. By the way, we should note here that this model practically does not work in the large-scale web applications. We think it is very important because without scalability we cannot think about “billions of connected devices”.
Application Interaction	The application interaction part is added in order to support development of simple M2M applications with only minor application specific data definitions: readings, observations and commands.	Application interactions build on the generic messaging and transaction functionality and offer capabilities considered sufficient for most simple application domains.
Messaging	The Message service capability feature offers message delivery with no message duplication. Messages may be unconfirmed, confirmed or transaction controlled.	The message modes supported are single Object messaging, Object group messaging, and any object messaging; (it can also be Selective object messaging). Think about this as Message Broker.
Event notification and presence	The notification service capability feature is more generic than handling only presence. It could give notifications on an object entering or leaving a specific group, reaching a certain location area, sensor readings outside a predefined band, an alarm, etc.	It is a generic form. So, for example, geo fencing should fall into this category too. The subscriber subscribes for events happening at the Target at a Registrar. The Registrar and the Target might be the same object. This configuration offers a publish/subscribe mechanism with no central point of failure.
Compensation	Fair and flexible compensation schemes between cooperating and competing parties are required to correlate resource consumption and cost, e.g. in order to avoid anomalous resource consumption and blocking of incentives for investments. The defined capability feature for micro-payment additionally allows charging for consumed network resources.	It is very similar, by the way, to Parlay’s offering for Charging API.
Sessions	In the context of OpenAPI a session shall be understood to represent the state of active communication between Connected Objects.	OpenAPI is REST based, so, the endpoints should be presented as some URI’s capable to accept (in this implementation) the basic commands GET, POST, PUT, DELETE (See an example below).

TABLE I. ETSI OPEN API CATEGORIES

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

A session example: requests execution of some function.

URI: http://{nodeId}/a/do

Method: POST

Request

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<appint-do-request xmlns="http://eurescom.eu/p1957/openm2m">
<requestor>9378f697-773e-4c8b-8c89-27d45ecc70c7</requestor>
<commands>
<command>command1</command>
<command>command2</command>
</commands>
<responders>9870f7b6-bc47-47df-b670-2227ac5aaa2d</responders>
<transaction-id>AEDF7D2C67BB4C7DB7615856868057C3</transaction-id>
</appint-do-request>
```

Response

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?><appint-do-response xmlns="http://eurescom.eu/p1957/openm2m">
<requestor>9378f697-773e-4c8b-8c89-27d45ecc70c7</requestor>
<timestamp>2010-04-30T14:12:34.796+02:00</timestamp>
<responders>9870f7b6-bc47-47df-b670-2227ac5aaa2d</responders>
<result>200</result>
</appint-do-response>
```

Note that because we are talking about server-side solution, there is no problem with so called sandbox restrictions. But it means of course, that such kind of request could not be provided right from the client side as many modern web applications do.

III. WEB INTENTS VS. OPEN API FROM ETSI

Let us start from the basic. Users use many different services on the web to handle their day to day tasks, developers use different services for various tasks. In other words, our environment consists of connected applications. And of course, all they expect their applications to be connected and to work together seamlessly.

It is almost impossible for developers to anticipate every new service and to integrate with every existing external service that their users prefer, and thus, they must choose to integrate with a few select APIs at great expense to the developer.

As per telecom experience, we can mention here the various attempts for unified API that started, probably, with Parlay. Despite a lot of efforts, Parlay API's actually increase the time for development. It is, by our opinion, the main reason for the Parlay's failure [9].

Web Intents solves this problem. Web Intents is a framework for client-side service discovery and inter-application communication. Services register their intention to be able to handle an action on the user's behalf. Applications request to start an action of a certain verb (for example share, edit, view, pick etc.) and the system will find the appropriate services for the user to use based on the user's preference. It is the basic [12].

Going to M2M applications it means that our potential devices will be able to present more integrated for the measurement visualization for example. The final goal of any M2M based application is to get (collect) measurements and perform some calculations (make some decisions) on the collected dataset. We can go either via low level API's or use (at least for majority of use cases) some integrated solutions. The advantages are obvious. We can seriously decrease the time for development.

Web Intents puts the user in control of service integrations and makes the developers life simple.

Here is the modified example for web intents integration for the hypothetical web intents example:

1. Register some intent upon loading our HTML document

```
document.addEventListener("DOMContentLoaded", function() {
    var regBtn = document.getElementById("register");
    regBtn.addEventListener("click", function() {
    window.navigator.register("http://webintents.org/m2m", undefined);
    }, false);
```

2. Start intent's activity

```
    var startButton = document.getElementById("startActivity");
    startButton.addEventListener("click", function() {
      var intent = new Intent();
      intent.action = "http://webintents.org/m2m";
        window.navigator.startActivity(intent);
    }, false);
```

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

3. Get measurements (note – in JSON rather than XML) and display them in our application

```
window.navigator.onActivity = function(data) {
  var output = document.getElementById("output");
  output.textContent = JSON.stringify(data);
 };
}, false);
```

Obviously, that it is much shorter than the long sequence of individual calls as per M2M Open API.

The key point here is *onActivity* callback that returns JSON (not XML!) formatted data. As per suggested M2M API we should perform several individual requests, parse XML responses for the each of them and only after that make some visualization. Additionally, web intents based approach is asynchronous by its nature, so, we don need to organize asynchronous calls by our own.

Also, Web Intents approach lets us bypass sandbox restrictions. In other words, developers can raise requests right from the end-user devices, rather than always call the server. The server-side only solution becomes bottleneck very fast. And vice-versa, client side based request let developers deploy new services very quickly. Why do not use the powerful browsers in the modern smart-phones? At the end of the day Parlay spec were born in the time of WAP and weak phones. Why do we ignore HTML5 browsers and JavaScript support in the modern phones?

IV. DATA PERSISTENCE

The next question we would like to discuss relating to the M2M API's is probably more discussion able. Shall we add some persistence API (at least on the form of generic interface)?

The reasons are obvious – save the development time. Again, we should keep in mind that we are talking about the particular domain – M2M. In the most cases, our business applications will deal with some metering data. As soon as we admit, that we are dealing with the measurements in the various forms we should make, as seems to us a natural conclusion – we need to save the data somewhere. It is very simple – we need to save data for the future processing.

So, the question is very easy – can we talk about M2M applications without talking about data persistence? Again, the key question is M2M. It is not abstract web API. We are talking about the well-defined domain.

As seems to us, even right now, before the putting some unified API in place, the term M2M almost always coexists with the term "cloud". And as we can see, almost always has been accompanied by the terms like automatic database logging, backup capabilities, etc.

So, maybe this question is more for the discussions or it even could be provocative in the some forms, but it is: why there is no reference API for persistence layer in the unified M2M API? It is possible in general to create data gathering API without even mentioning data persistence?

V. NEW SIGNALING DEMAND

Eventually, billions of devices — such as sensors, consumer electronic devices, smart phones, PDAs and computers — will generate billions of M2M transactions. For example: Price information will be pushed to smart meters in a demand-response system. Push notifications will be sent to connected devices, letting a client application know about new information available in the network. The scale of these transactions will go beyond anything today's largest network operators have experienced. Signaling traffic will be the primary bottleneck as M2M communications increase. Alcatel-Lucent Bell Labs traffic modeling studies support this by comparing network capacity against projected traffic demand across multiple dimensions (such as signaling processing load on the radio network controller, air-interface access channel capacity, data volume and memory requirement for maintaining session contexts). The limiting factor is likely to be the number of session set-ups and tear-downs. For the specific traffic model and network deployment considered in the study, it is seen that up to 67 percent of computing resources in the radio network controller is consumed by M2M applications [5].

How much of the traffic sent is network overhead? As an analysis carried on by A. Sorrevad [13] shows for ZigBee solution, a node is sending at least 40 Mbytes per year with the purpose of maintaining the network and polling for new data. The trigger data traffic for a year is much less - around1-10 Mbytes. Thus, we see that the relationship between network and trigger traffic can range between 40:1 to 4:1 in a ZigBee solution that is following the home automation specification.

The traffic sent when maintaining a 6LoWPAN network is application specific. The relationship between network and trigger traffic can then be in the range 2:1 to 5:1.

Why do we think it is a place for traffic talk? Because again, it is not clear completely how can we support transactional API's (as per ETSI draft [8]), without the dealing with the increased traffic. Simply – in our transactions we need the confirmation that device is alive, that operation has been performed, etc. All this is signaling traffic. Actually, this may lead to next provocative questions: do we really need transactional calls for all use cases? For example, the modern large-scale web applications (e.g., social networks) are not transactional internally.

VI. CONCLUSION

This article describes the current state for the open unified M2M API. Article proposes some new additions – web intents

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

as add-on for the more traditional REST approach. The main goal for our suggestions is the simplifying the development phases for M2M applications. The key advantages are JSON versus XML, asynchronous communications and integrated calls. Also we would like to point attention to the couple of important questions that are not covered yet: data persistence and signaling traffic.

VII. ACKNOWLEDGEMENT

The paper is financed from ERDF's project SATTEH (No. 2010/0189/2DP/2.1.1.2.0/10/APIA/VIAA/019) being implemented in Engineering Research Institute «Ventspils International Radio Astronomy Centre» of Ventspils University College (VIRAC).

VIII. REFERENCES

[1] M. Chen, J. Wan, and F. Li Machine-to-machine communications: architectures, standards, and applications. KSII T Internet Inf 6(2): pp. 471–489, 2012

[2] Standartisation mandate to CEN, CENELEC and ETSI in the field of measuring instruments for the developing of an open architecture for utility meters involving communication protocols enabling interoperability, European Commission, M/441, 2009.

[3] Managed and Cloud Services Insight Group Machine-to-Machine and Cloud Services: http://www.cisco.com/en/US/solutions/collateral/ns341/ns849/ns1098/white_paper_c11-663879.html. Retrieved: Mar, 2012

[4] H. Park, B. Kim, Y. Ko, and D. Lee "InterX: A service interoperability gateway for heterogeneous smart objects" in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference, 21-25 March 2011, pp. 233 – 238.

[5] H. Viswanathan,, "Getting Ready for M2M Traffic Growth" http://www2.alcatel-lucent.com/blogs/techzine/2011/getting-ready-for-m2m-traffic-growth/ Retrieved: Mar, 2012

[6] EURESCOM project P1957, Open API for M2M applications, http://www.eurescom.de/public/projects/P1900-series/P1957/. Retrieved: Mar, 2012

[7] Sensei project http://www.sensei-project.eu/. Retrieved: Feb, 2012

[8] Draft ETSI TS 102 690 V0.13.3 (2011-07) Technical Specification.

[9] J. Yim, Y. Choi, and B. Lee "Third Party Call Control in IMS using Parlay Web Service Gateway Advanced Communication Technology", 2006. ICACT 2006. The 8th International Conference Issue Date,: 20-22 Feb. 2006, pp. 221 – 224.

[10] I. Grønbæk., "Architecture for the Internet of Things (IoT): API and interconnect", The Second International Conference on Sensor Technologies and Applications, IEEE August 2008, DOI 10.1109/SENSORCOMM.2008.20, 809.

[11] I. Grønbæk and K. Ostendorf "Open API for M2M applications" In: ETSI M2M Workshop Oct. 2010.

[12] Web Intents http://webintents.org/ Retrieved: Mar, 2012

[13] A. Sorrevad "M2M Traffic Characteristics", KTH Information and Communication Technology, Master of Science Thesis, Stockholm, Sweden, 2009 TRITA-ICT-EX-2009:212 http://web.it.kth.se/~maguire/DEGREE-PROJECT-REPORTS/091201-Anders_Orrevad-with-cover.pdf Retrieved: Mar, 2012

[14] Axeda platform http://developer.axeda.com Retrieved: Feb, 2012

[15] 3GPP TS 22.368 V11.0.1, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Service requirements for Machine-Type Communications (MTC); Stage 1, (Release 11)

[16] J. Han, A. Vu, J. Kim, J. Jeon, S. Lee, and Y. Kim; "The fundamental functions and interfaces for the ITU-T USN middleware components", Information and Communication Technology Convergence (ICTC), 2010 International Conference, 17-19 Nov. 2010, pp.: 226 – 231. Print ISBN: 978-1-4244-9806-2.

[17] K. Chang, A. Soong, M. Tseng, and X.Zhixian Global Wireless Machine-to-Machine Standardization. Internet Computing, IEEE, March-April 2011, Vol.: 15, Issue: 2, pp. 64 - 69

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Fast Retrial and Dynamic Access Control Algorithm for LTE-Advanced Based M2M Network

Zhefeng Jiang, Xiaofeng Zhong
Dept. Electronic Engineering, Tsinghua National Laboratory for Information Sciences and Technology
Tsinghua University
Beijing, P.R.China
Email:jiangzf09@mails.tsinghua.edu.cn, zhongxf@tsinghua.edu.cn

***Abstract*—Currently, increasing number of devices are connected to networks. Hence, M2M (Machine to Machine) communication, especially LTE/LTE-Advanced based M2M communication, is attracting more and more interests from the telecommunication industry. However, as the current LTE (Long Term evolution) system is designed for Human to Human system, it may be unable to support the massive M2M devices. This paper proposes a fast retrial and dynamic random access algorithm for LTE-Advanced network. It drastically reduces the delay of access comparing the back-off algorithm. In the meantime, it successfully prevents the system from severe congestion, which is inevitable in back-off algorithm when the arrival rate of random access is very high. To make the dynamic control of random access feasible in practical scenario, an estimation algorithm of the access arrival rate is also proposed in this paper. Simulation results reveal that the algorithm is able to provide better delay performance and greater throughput comparing with the back-off schemes defined in the LTE network.**

***Index Terms*—M2M, Random Access, Dynamic control, LTE-Advance**

I. INTRODUCTION

M2M (machine to machine) communication is considered to be a new killer application for the next generation communication system, such as LTE-Advanced network. However, current communication systems are designed for Human to Human communication, while M2M applications are characterized by some unique features such as massive nodes, extremely high frequency of accessing [1], [2] and so on. Hence, in order to better support M2M applications, it is necessary to introduce some specific optimization for M2M communication to the LTE-Advanced system. One of the most important issue is designing an effective medium access scheme to handle the high frequency of access of the massive MTC (machine-type communications) devices.

Random access is popular for medium access control. Moreover, time slotted random access is adopted in 2G (2rd Generation), 3G (3rd Generation) and LTE (Long Term Evolution) for initiating uplink access because of its channel efficiency in licensed channels [3]. Back-off algorithm has been adopted in LTE to alleviate grave congestion following random access collision. Sharma et al. [4] studied the performance of back-off based random access in IEEE 802.11 DCF. Nevertheless, Back-off scheme brings about great time delay and fails to deal with circumstances with extremely high arrival rate of accesses, especially when there are massive nodes. Aldous et al. [5] evaluated the ultimate instability of exponential back-off protocol with transmission control for random access. Rivest et al. [6] and Hauksson et al. [7] proposed algorithms adopting other dynamic medium access control methods to improve the delay and loss performance of the system. However, they can only be used in single channel. Choi et al. [8] proposed a multichannel random access with fast retrial. It can successfully limit the time delay in random access. However, it is unable to sustain stable even when the arrival rate of access is not very high.

In this paper, a fast retrial and dynamic access control algorithm is proposed to deal with the congestion in multichannel random access under extremely high arrival rate of access. It is able to achieve a comparatively low delay, in addition to effectively utilizing the channels. Furthermore, it works well even when the arrival rate of access is higher than the limitation of time slotted aloha scheme e^{-1} [9].

The paper is organized as follows: Section II introduces the system model, and describes the proposed algorithm in detail. In Section III, the algorithm's performance is analyzed. Section IV shows simulation results. Finally, the paper is concluded in section V.

II. SYSTEM MODEL AND OPTIMIZED ALGORITHM

A. Uplink random access

Consider that in a LTE-Advance system, there are numerous MTC nodes. Each node needs to conduct a contention-based random access procedure to obtain an uplink channel. The physical resource of random access in LTE-Advance includes preambles and random access opportunities. In LTE-Advance, each cell is allocated with 64 preambles, and some preambles are assigned to non-contention-based random access. For cells below 1.5km radii, all 64 preambles are orthogonal to each other as they are derived from single root Zadoff-Chu sequence. In larger cells, though 64 preambles are not perfectly orthogonal to each other as they are derived from multiple root sequences, the cross-correlation is low [10]. Hence, we assume that all preamble are orthogonal to each other and one preamble is denoted as a logical channel in this paper. According to the current LTE-Advance network definition, time slotted aloha is adopted in random access. In which each node can send a random access request in the dedicated time slot (random access opportunity).

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Assuming that there are N orthogonal preambles, that is, there are N parallel logical channels (mentioned as channels in this paper) in one random access opportunity (mentioned as slot). In each slot, if more than one MTC nodes have sent the same preamble, a collision will happen. Fig. 1 illustrates a abstract system with 4 logical channels and 3 random accesses in the first slot.

The contention-based random access procedure in LTE-Advance is outlined in Fig. 2, and it is further described below. Readers can get more information about the contention-based random access from [3], [11].

1) Random access preamble. Each ue randomly select a preamble from the contention-based-group and transmit it in a nearby random access opportunity.
2) Random access response. The eNodeB correlates all possible preambles in each random access opportunity with the received preamble. With the detected preambles, the eNodeB assigns uplink resources related MTC nodes and broadcasts the the information.
3) Scheduled Transmission. MTC nodes transmits unique identity with the allocated uplink resource.
4) Contention resolution. In step 3, more than one MTC nodes which had sent the same preamble may response. In this case, the eNodeB is unable to decode the identities from these nodes. So these nodes will not receive the notification of the reception step 3 in the dedicated time window, and they will go to step 1.

In the back-off algorithm of current LTE, before conducting step 1, each node needs to wait for a randomly determined number of slots (back-off time). Where the back-off time follows Uniform distribution on [0, maximum back-off slot]. If the random access procedure is unsuccessful after maximum retrial times of step 1, the random access will be abandoned.

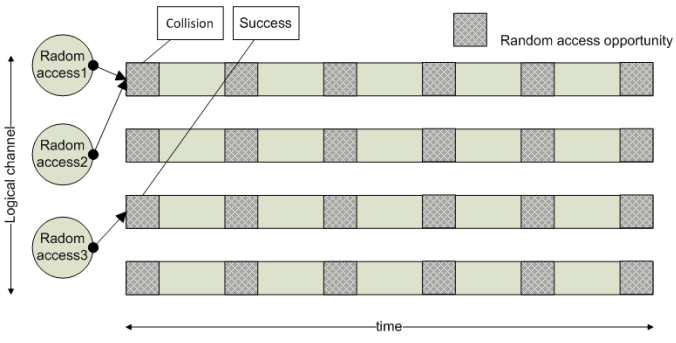


Fig. 1. A system with 4 logical channel

B. Proposed algorithm

In the proposed algorithm, MTC nodes which have suffered from collision will send a preamble immediately in the next random access opportunity. Before every random access opportunity, the eNodeB broadcasts an access rate. For all MTC nodes that need to conduct a random access, they send a preamble at a rate of α. The parameter α is dynamically determined by the eNodeB and timely broadcasted by it. If a MTC node is about to send a access request, it randomly choose one channel from all channels with equal possibility. If a collision happens, it will retry in the next slot with a randomly chosen channel. The flow diagram of the proposed algorithm (first random access requests) is shown in Fig. 2:

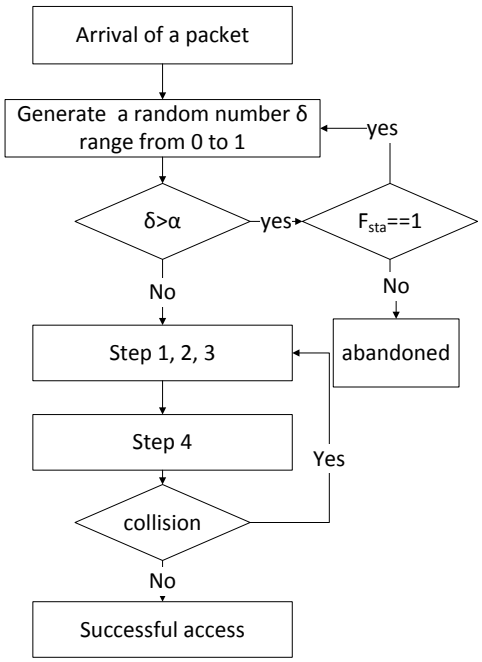


Fig. 2. flow diagram of the proposed algorithm

Now, we define the following notations. They will be used in parameter determinations and performance evaluation.

N	number of random access channels
λ	New arrivals of random access
λ'	combined arrivals of random access, including new and time-domain backlogged arrivals
M	number of retrial access in current random access opportunity.
α	the rate of sending a preamble in random access
p_{uu}	The possibility of a channel is used
p_{nc}	The possibility of a channel has no collision
p_{uu}	The possibility of a channel is not used
n_{nc}	Number of logical channels with no collision in the current slot
n_{uu}	Number of unused channels in the current slot
n_{ac}	number of logical channel with a successful random access in the current slot
p_{su}	the possibility that an access and be succeed at last
p_{ac}	the possibility that an access is accepted in current slot
F_{sta}	When it equals to 1, the system is considered to be stable. When it equals to 0, it is considered to be unstable.
W	contention resolution window

C. The determination of α and F_{sta}

According to the flow diagram of the algorithm, each MTC node needs to receive α and F_{sta} from the eNodeB before each slot.

Proposition 1: Assuming that the combined arrivals in a time slot follow Poisson distribution with a mean λ'. There are M retrials and N logical channel$(N > 1, 0 \leq M < N)$. In the proposed algorithm, the optimal α follows:

$$\alpha = \max(0, \min((N - M - 1)\frac{N}{(N-1)\lambda'}, 1)) \qquad (1)$$

Proof: Without loss of generality, in the $i^{th}, (i = 1, 2, 3 \cdots N)$ channel, the combined arrivals follow Poisson with mean $\frac{\lambda'}{N}$. The retrial accesses n_i^r, and combined arrivals n_i^c in the i^{th} channel follow the following distribution:

$$P(n_i^r = k) = C_M^k (\frac{1}{N})^k (\frac{N-1}{N})^{M-k}, k = 1, 2, \cdots M \quad (2)$$

$$P(n_i^c = j) = \frac{(\frac{\alpha\lambda'}{N})^j}{j!} e^{-\frac{\alpha\lambda'}{N}}, 0 < \alpha \leq 1 \quad (3)$$

The accessing in the $i^{(th)}$ channel will succeed if, in the same slot, the channel is only used by one MTC device. Hence, the possibility of success random access in the $i^{(th)}$ channel $p_{i,ac}$ follows:

$$\begin{aligned} & p_{i,ac} = P(n_i^r + n_i^c = 1) \\ & = P(n_i^r = 0, n_i^c = 1) + P(n_i^r = 1, n_i^c = 0) \\ & = C_M^1 (\tfrac{1}{N})(\tfrac{N-1}{N})^{M-1} \cdot e^{-\frac{\alpha\lambda'}{N}} + C_M^0 (\tfrac{N-1}{N})^M \cdot \tfrac{\alpha\lambda'}{N} e^{-\frac{\alpha\lambda'}{N}} \end{aligned} \quad (4)$$

$p_{i,ac}$ achieves the maximum value when α is determined as equation (1).

Without loss of generality, it can be applied to other channels. Hence, in the proposed algorithm, equation (1) denotes the optimal α of the system.

■

Therefore, the eNodeB can determine the optimized α with λ' and M. The estimating of these two parameters is presented as follow:

Estimation of the new arrival rate λ***:*** When the system is stable, the leaving rate(throughput) of the system equals to the arrival rate of new access. Hence, we can reach an reliable estimation of the arrival rate of new accesses λ with the leaving rate of the system. Denote F_k as the number of successful accesses in k^{th} slot. We have:

$$\tilde{\lambda}_k = \begin{cases} \sum_{i=1}^{L} \frac{F_{k-i}}{L} & k - L > j \\ \sum_{i=j}^{k-1} \frac{F_i}{k-j} & k - L \leq j \end{cases} \quad (5)$$

where $j - 1$ is nearest slot that is considered unstable.

When the arrival rate of random access λ is beyond the maximum throughput, the system must be unstable. Hence, we have to find another reliable method to estimate the λ.

Furthermore, in this circumstance, it is meaningless to backlog the access that has not been sent when it arrives, because it is impossible to handle all random access in this condition. So any arrived access will be sent immediately in the nearby slot or be abandoned. Hence, we have $\lambda \equiv \lambda'$.

The unused possibility of a channel follows:

$$p_{uu} = P(n_i^r + n_i^c = 0) = C_M^0 (\frac{N-1}{N})^M \cdot e^{-\frac{\alpha\lambda}{N}} \quad (6)$$

So, we have

$$\lambda = -\frac{N}{\alpha} \ln(p_{uu} \cdot (\frac{N}{N-1})^M) \quad (7)$$

We can estimate the arrival rate of random access with:

$$\lambda_k = -\frac{N}{\alpha} \ln(\tilde{p}_{uu,k} \cdot (\frac{N}{N-1})^{\tilde{M}_k}) \quad (8)$$

where $\tilde{p}_{uu,k} = \begin{cases} \sum_{i=1}^{L} \frac{n_{uu,k-i}}{LN} & k - L > j' \\ \sum_{i=j'}^{k-1} \frac{n_{uu,i}}{k-j'} & k - L \leq j' \end{cases}$, $\tilde{M}_k = \begin{cases} \sum_{i=1}^{L} \frac{M_{k-i}}{L} & k - L > j' \\ \sum_{i=j'}^{k-1} \frac{M_{j'}}{k-j'} & k - L \leq j' \end{cases}$, and $j' - 1$ is the nearest slot that is considered stable.

Estimation of the number of retrial random access M***:*** We assume that the eNodeB is able to detect all collisions in step 3, and the retrial slot of the corresponding nodes can be determined with the contention resolution window W. However, it is impossible for the eNodeB to identify the number of accesses in a collided channel. Hence, it is necessary to propose a reliable method to estimate it. Assume that S scheduled transmission have been send during the k^{th} slot, we have:

$$p_{nc} = \begin{cases} 1 & s = 0, 1 \\ C_S^0 (\frac{N-1}{N})^S + C_S^1 \frac{1}{N} (\frac{N-1}{N})^{S-1} & s > 1 \end{cases} \quad (9)$$

We estimate the possibility of no collision happens in certain channel with

$$\tilde{P}_{nc} = n_{accept}/N \quad (10)$$

We have:

$$\tilde{M} = \tilde{S} - n_{ac} \quad (11)$$

Finally, in the $(K + W)_{th}$ slot, $\tilde{S}$ can be calculated with $\tilde{P}_{nc}$ by looking-up a table established according to equation (9), as there is no analytical solution for equation(9), and (9) is a monotone function about S.

Evaluating the combined arrivals λ'***:*** The parameter α is dynamically adjusted in each slot. If we denote λ'_k as the combined arrivals in k^{th} slot, we have:

$$\lambda'_k = \begin{cases} \lambda'_{k-1}(1 - \alpha_{k-1}) + \lambda_k & F_{sta,k} = 1, k > 1 \\ \lambda_k & F_{sta,k} = 0, k > 1 \\ 0 & k = 1 \end{cases} \quad (12)$$

The determination of F_{sta}***:*** When λ is higher than the maximum throughput, estimating λ with the leaving rate must result in deviation of the estimation. Hence, it is necessary to identify the unstable state in time. In this paper, we use the non-collision rate of all channels as the indicator. Besides, to limit the delay of succeeded access, the system is also considered to be unstable when $\alpha < \alpha_0$ in this paper.

Assume that S accesses have been sent in k^{th} slot, the Maximum throughput of the system can be reached when $S = 1/\ln(\frac{N}{N-1})$. In this circumstance, the non-collision rate of the channels follows:

$$\begin{aligned} & P'_{nc} = C_S^0 (\tfrac{N-1}{N})^S + C_S^1 \tfrac{1}{N} (\tfrac{N-1}{N})^{S-1} \\ & = \left(1 + \tfrac{1}{(N-1)\ln(\frac{N}{N-1})}\right) e^{-1} \end{aligned} \quad (13)$$

If in the $(k-1)^{th}$ slot, the system is considered to be stable, in the $(k)^{th}$ slot, the system state is determined as follow:

$$F_{sta,k} = \begin{cases} 0 & \tilde{P} < P'_{nc}, or\ \alpha_k - 1 < \alpha_0 \\ 1 & otherwise \end{cases} \quad (14)$$

where $\tilde{P} = \frac{\sum_{i=1}^{L} n_{nc,k-i}}{N}$

When the system is considered to be unstable and the new arrival rate λ becomes small enough to make the system considered stable, it is necessary to shift the system state from unstable to stable in time. Denote $\overline{\alpha} = \frac{\sum_{i=1}^{i=L} \alpha_{k-i}}{L}$. If the system is considered to be unstable in $(k-1)^{th}$ slot, the state of the system in k^{th} is determined as follow:

$$F_{sta,k} = \begin{cases} 1 & \overline{\alpha} < 1 \\ 0 & otherwise \end{cases} \quad (15)$$

Besides, $F_{sta,1}$ is set to be 0.

III. Performance evaluation

According to the results of proposition 1, the system achieves the maximum throughput when

$$\alpha\lambda' = (N - M - 1)\frac{N}{(N-1)}, (N > 1, 0 \le M < N) \quad (16)$$

So, the maximum throughput of the system can be denoted by:

$$\begin{aligned} &P(n_i^r + n_i^c = 1) \\ &= C_M^1(\tfrac{1}{N})(\tfrac{N-1}{N})^{M-1} \cdot e^{-\frac{\alpha\lambda'}{N}} + C_M^0(\tfrac{N-1}{N})^M \cdot \tfrac{\alpha\lambda'}{N} e^{-\frac{\alpha\lambda'}{N}} \\ &= C_M^1(\tfrac{1}{N})(\tfrac{N-1}{N})^{M-1} \cdot e^{-\frac{1}{N}(N-M-1)\frac{N}{(N-1)}} \\ &+ C_M^0(\tfrac{N-1}{N})^M \cdot \tfrac{1}{N}(N-M-1)\tfrac{N}{(N-1)} e^{-\frac{1}{N}(N-M-1)\frac{N}{(N-1)}} \\ &= (1 - \tfrac{1}{N})^M \cdot e^{-\frac{N-M-1}{N-1}} \end{aligned} \quad (17)$$

when $N > 1, 0 \le M < N$, we have:

$$e^{-1} \le (1 - \frac{1}{N})^M \cdot e^{-\frac{N-M-1}{N-1}} \le \frac{1}{2} \quad (18)$$

Hence, the Maximum throughput per slot of the system is:

$$N \cdot (e^{-1} + \delta), 0 \le \delta \le \frac{1}{2} - e^{-1} \quad (19)$$

The performance of the system is evaluated separately according to whether the system is considered to be stable.

When the system is considered to be stable, the possibility of successful random access follows:

$$p_{ac} = P(n^r + n^f = 1) * N/(M + \alpha\lambda') \quad (20)$$

where α is determined by equation (1), and $\alpha\lambda' \le (N - M - 1)\frac{N}{(N-1)}$, for α may be smaller than 1 when λ' is very low.

Hence, we have:

$$\begin{aligned} p_{accept} &= P(n^r + n^f = 1) * N/(M + \alpha\lambda') \\ &= \frac{N(C_M^1(\frac{1}{N})(\frac{N-1}{N})^{M-1} \cdot e^{-\frac{\alpha\lambda'}{N}} + C_M^0(\frac{N-1}{N})^M \cdot \frac{\alpha\lambda'}{N} e^{-\frac{\alpha\lambda'}{N}}))}{M + \alpha\lambda'} \\ &\ge \frac{N(N-1/N)^M e^{-\frac{N-M-1}{N-1}}}{N + M - MN/N - 1} \\ &\ge 1/e \end{aligned} \quad (21)$$

When the system is stable, all accesses would retry until they are successfully accepted by the eNodeB. Hence, we have $p_{success} = 1$, and the throughput equals to arrival rate λ.

When the system is considered unstable, we have $\lambda = \lambda'$.

With the results of proposition1, the system can achieve the maximum throughput with $\alpha = (N - M - 1)\frac{N}{(N-1)\lambda'}$.

Therefore, the throughput of the system follows $N \cdot (e^{-1} + \delta), 0 \le \delta \le \frac{1}{2} - e^{-1}$.

The success rate of access follows:

$$p_{su} = \frac{(e^{-1} + \delta)}{\lambda} \ge \frac{1}{e\lambda} \quad (22)$$

IV. Simulations

In this section, we use Matlab-based simulation to compare the proposed algorithm with back-off schedule, which is adopted in the current LTE system.

Simulation parameters are shown in table 1. The new arrival rate of new random access has already been normalized. That is, λ denotes the mean number of new random access per channel per slot.

TABLE I. parameter settings

Parameters	Back-off (1)	Back-off (2)	Proposed (1)	Proposed (2)
Number of nodes	1000	1000	1000	1000
Number of preambles	16	16	16	16
Maximum back-off slot	20	40	-	-
Maximum retrial times	5	5	-	-
α_0	-	-	0.2	0.4

We compare the two algorithms with the following indices:

Average delay:	The mean delay of successful accesses
Throughput:	The mean successful access in a slot
Success ratio:	the ratio of accesses that can finally successfully accepted

Fig. 3 shows that the proposed algorithm has a much better delay performance than the back-off algorithm. Fig. 4 shows that the throughput of the proposed algorithm keeps stable when λ is higher than $1/e$. But the throughput in back-off algorithm decreases rapidly when λ is higher than $1/e$. After all, the system fail to keep stable when the arrival rate is higher than the Maximum throughput of the slotted aloha. It is shown that the maximum throughput of the proposed algorithm is higher than that of time slotted Aloha with Poisson arrivals, which is $16/e$ with 16 channels. Besides, the line N/e corresponds to the maximum throughput of slotted aloha with N channels. It is because that, in the proposed algorithm, the number of accesses in each slot no longer

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

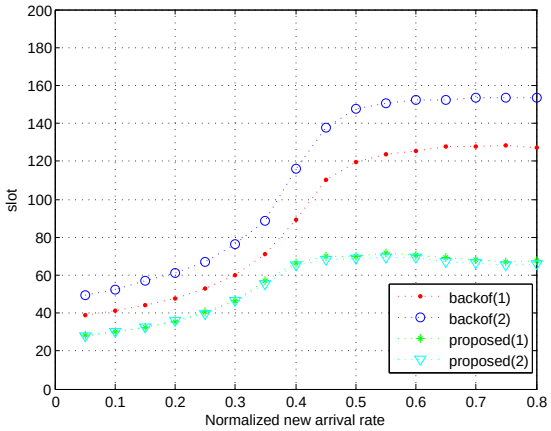


Fig. 3. Average delay

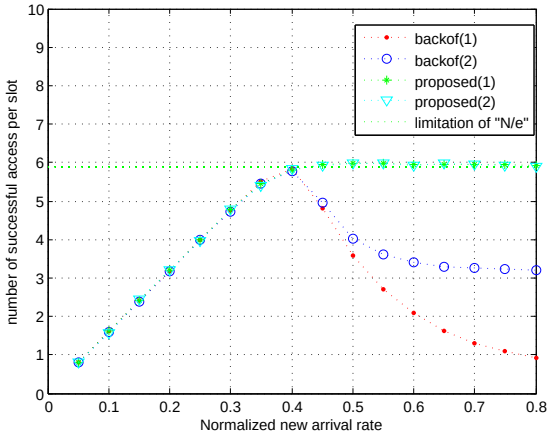


Fig. 4. Throughput

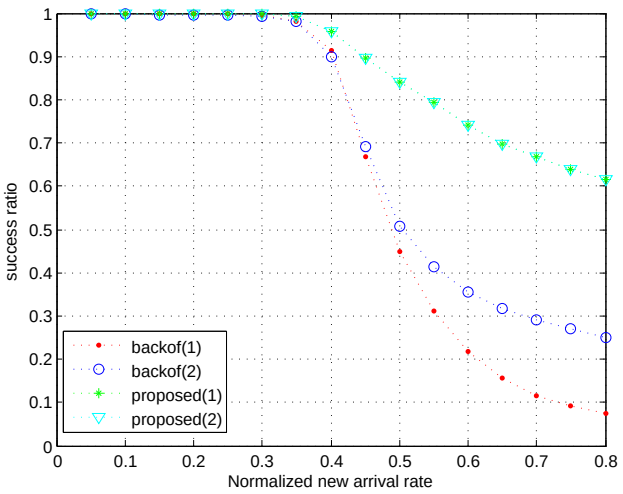


Fig. 5. Success ratio

follows Poisson distribution owing to the fast retrial scheme. Fig. 5 shows that the success ratio of the proposed algorithm keeps stable when λ is higher than $1/e$, while the throughput of back-off algorithm decreases rapidly when λ is higher than $1/e$.

V. Conclusion

Through dynamic control of the contention-based random access, and the fast retrial of collided accesses depends on the reliable estimation methods, the proposed algorithm is able to guarantee the reliability of the system under extremely high rate of access, as shown in analysis and simulations. Besides, the maximum utilization of the channel is close to or even above the extreme utilization of time slotted Aloha system with Poisson arrivals. Moreover, the delay of random access is limited. In conclusion, the proposed algorithm can well serve the M2M applications in LTE-Advance, which are featured by extremely high rate of accesses.

Acknowledgment

This work is supported by National S&T Major Project (2011ZX03005-003-01), and Chinese 973 Program (2012CB316006).

References

[1] "Service requirements for machine-type communications," tech. rep., 3GPP TS 22.368 V11.2.0, June, 2011.

[2] "System improvements for machine-type communications," tech. rep., 3GPP TR 23.888 V1.2.0, April, 2011.

[3] "Medium access control (mac) protocol specification," tech. rep., 3GPP TS 36.321 V10.2.0, June, 2011.

[4] G. Sharma, A. Ganesh, and P. Key, "Performance analysis of contention based medium access control protocols," *Information Theory, IEEE Transactions on*, vol. 55, pp. 1665 –1682, April 2009.

[5] D. Aldous, "Ultimate instability of exponential back-off protocol for acknowledgment-based transmission control of random access communication channels," *Information Theory, IEEE Transactions on*, vol. 33, pp. 219 – 223, Mar 1987.

[6] R. Rivest, "Network control by bayesian broadcast," *Information Theory, IEEE Transactions on*, vol. 33, pp. 323 – 328, May 1987.

[7] G. Hauksson and M. Alanyali, "Wireless medium access via adaptive backoff: Delay and loss minimization," in *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, pp. 1777 –1785, April 2008.

[8] Y.-J. Choi, S. Park, and S. Bahk, "Multichannel random access in ofdma wireless networks," *Selected Areas in Communications, IEEE Journal on*, vol. 24, pp. 603 –613, March 2006.

[9] R. G. Dimitri Bertsekas, *Data networks,2nd ed.* Englewood Cliffs, N.J. : Prentice Halls, 1992.

[10] M. Amirijoo, P. Frenger, F. Gunnarsson, J. Moe, and K. Zetterberg, "On self-optimization of the random access procedure in 3g long term evolution," pp. 177 –184, June 2009.

[11] "Evolved universal terrestrial radio access (e-utra) and evolved universal terrestrial radio access network," tech. rep., 3GPP TS 36.300 V10.5.0, September, 2011.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Rotated Constellations with Scaled Factor for High-Rate Full-Diversity STBC of 2 and 4 Antennas

Yier Yan[1] [2], Jun Li[2], Tae Chol Shin[2], Moon Ho Lee[2]
School of Mechanical and Electric Engineering, Guangzhou University[1], Guangzhou, 51006, PR China
Dep. Electric and Computer Engineering Chonbuk National University[2]
664-14 Deokjin-dong, Jeonju 561-756, Republic of Korea
Email: year0080@gzhu.edu.cn, Junli0502@jbnu.ac.kr, tcshin@naver.com, moonho@chonbuk.ac.kr

***Abstract*—We design a new** $rate-5/4$ **full-diversity orthogonal space-time block code (STBC) transmission scheme for QPSK with 2 transmitting antennas (TX) by one modified QPSK constellation with rotated and scaled factor by maximizing the CGD (coding gain distance) from the set of quaternions used in the Alamouti code. A low-complexity maximum likelihood (ML) decoding algorithm has been proposed that not only provide good FER (frame error rate) as same as** $rate-1$ **but also increase the transmitted rate approach the** $rate-9/8$ **without additional bandwidth. Finally, we extend the design to the case for 4 TX with low complexity by enlarging the set of Quasi-Orthogonal STBC proposed in without power scaling. Extension to general M-PSK constellation is also straightforward is the simulation results. The simulation result shows that the proposed scheme provides a better Channel Throughput performance than the original work.**

Keywords-STBC; ML; Constellation; High-rate; MPSK.

I. INTRODUCTION

SPACE–TIME trellis codes have been introduced in [1] to provide an improved error performance for wireless systems with multiple transmit antennas. The authors have shown that such codes can provide full diversity gain as well as additional signal-to-noise ratio (SNR) advantage that they call the coding gain. Code design rules for achieving full diversity are also provided. Using these design rules, examples of codes with full diversity as well as some coding gain constructed are not necessarily optimal. Since there is no general rule for designing codes that provide diversity as well as coding gain, it is unclear how to design new codes for different number of states or different rates. Also, it is not clear how to improve the performance of the codes, i.e., how to maximize the coding gain. There have been many efforts to improve the performance of the original space–time trellis codes [2]–[5]. While very interesting codes have been proposed in the literature, the coding gain improvements are marginal for one receive antenna.

In [8], the authors have proposed a class of STBCs for a high rate transmission scheme by exploiting the inherent algebraic structure for 2 and 4 transmit antennas. Compared with Alamouti and Jafarkhani schemes, the diversity gain is attenuated due to rotating or scaling one transmit matrix selected from the class.

In this work, we provide a new structure for space–time trellis codes that guarantees full diversity and provides opportunity to maximize the coding gain. We also provide a systematic method to maximize the coding gain for a given rate (rate>1), constellation, and number of states. The main idea of the proposed scheme is to employ two different signal constellations, and each constellation contains the information of the transmitted bit. In order to achieve the best system performance, the design criterion is also introduced in this paper to maximize the CGD to get the optimum rotated angle and scaled factor. The simulation is also given to demonstrate the reliability of the proposed transmission scheme.

The organization of the paper is as follows. In Section II, we briefly described the transmission scheme. Then, the proposed constellation is briefly introduced in Section III. In Section IV, the proposed algorithm is also produced to improve the system performance. Finally, simulation results are used to demonstrate the proposed scheme in Section V.

II. TRANSMISSION SCHEME

The objective in this paper is to design a new transmission scheme for high-rate (> 1) space-time block codes (STBC) by exploiting the QPSK constellation structure in existing orthogonal designs based on quaternions for 2 transmitting antennas [1] and quasi-orthogonal designs for 4 transmitting antennas [5]. The simplest example of a complex orthogonal design is the 2×2 code

$$G(x_1,x_2)=\begin{bmatrix} x_1 & x_2 \\ -x_2^* & x_1^* \end{bmatrix} \quad (1)$$

proposed by Alamouti [1], where $(\cdot)^*$ denotes the complex conjugate transpose and $x_i \in s, i=1,2$ and $s \in \left\{1/\sqrt{2}\cdot(\pm 1 \pm i)\right\}$. This code achieves rate-1 at full diversity and enjoys low-complexity ML decoding by employing matched filtering. The main idea of this paper is to enlarge the transmitted signaling set with maximized CGD [1] and full diversity. Then, we enlarge the signaling set based on the STBC Matrix $G(x_1,x_2)$. Two different QPSK constellations

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

proposed by rotating and scaling the conventional constellation are considered. The new constellation can be constructed by multiplying a suitable angle $\exp(j\theta)$ and an amplitude λ. The mathematical equation considered in this paper to modify the QPSK is given by $s' = s \cdot \lambda \exp(j\theta),\ |\lambda| \le 1, |\theta| \le \pi,$ and $s \in \{1/\sqrt{2} \cdot (\pm 1 \pm i)\}$. The two different constellations are given in the Fig. 1, in which the outer circle can be expressed the conventional QPSK constellation belonged to signal set.1 s and the inner circle can be expressed as the modified signal constellation s'. A very important criterion will be greatly discussed and calculated in this paper derived in [2] $\det((C_i - C_j)'(C_i - C_j))$, and the difference of two codes can be defined as $C_i - C_j = D_{ij}$, here the entrance of G_i belongs to constellation *s*, and each the entrance of $G_j^{'}$ belongs to s', then we briefly prove that $\det(D_{ij}^{'} D_{ij})$ be full diversity,

$$\det(D_{ij}^{'} D_{ij}) = \det\left(\begin{pmatrix} x_1(1-\lambda e^{j\theta}) & x_2(1-\lambda e^{j\theta}) \\ -x_2^*(1-\lambda e^{-j\theta}) & x_1^*(1-\lambda e^{-j\theta}) \end{pmatrix}^2\right)$$

$$= \det\left(\begin{pmatrix} x_1(1-\lambda e^{j\theta}) & x_2(1-\lambda e^{j\theta}) \\ -x_2^*(1-\lambda e^{-j\theta}) & x_1^*(1-\lambda e^{-j\theta}) \end{pmatrix}\right)^2 \quad (2)$$

while in terms of suitable angle θ and amplitude λ, the $C_i - C_j = D_{ij}$ is full diversity.

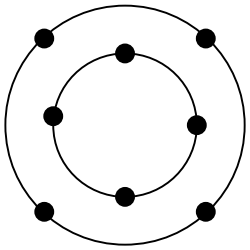


Fig. 1 Conventional and proposed constellations

Then, G_0 and G_1 are given here, the entrances of G_0 are selected from conventional QPSK constellation and the entrances of G_1 are selected from proposed constellation for the transmission scheme. We will use the expanded set S to construct new high-rate (>1) full diversity space-time block code with low decoding complexity and optimized available coding gain.

$$G_0(x_1, x_2, b_0 = 0) = \begin{bmatrix} x_1 & x_2 \\ -x_2^* & x_1^* \end{bmatrix},$$
$$G_1(x_1, x_2, b_0 = 1) = \begin{bmatrix} x_1\lambda e^{j\theta} & x_2\lambda e^{j\theta} \\ -x_2^*\lambda e^{j\theta} & x_1^*\lambda e^{j\theta} \end{bmatrix}, \quad (3)$$

We will use the expanded set S to construct new high-rate (> 1) and full diversity space-time block code with low complexity decoding and optimized available coding gain. The space-time code G_0 is selected to transmit symbols while the additional bit b_0 is 0 to be transmitted. With this idea, G_1 is selected for transmitting $b_0 = 1$. Then, each transmission frame consists of two symbols by employing two different constellations with 5 information bits. Without additional system source, the transmission can be improved according such transmission scheme.

III. THE PROPOSED CONSTELLATION FOR TWO ANTENNAS

A. Design Criterions for the Proposed Constellation

Consider two distinct codewords $C_i, C_j \in C'$. In order to ensure full spatial diversity, the codeword difference matrix $C_i - C_j = D_{ij}$ between any two distinct codewords in the extended set C must be full rank [4]. When both codewords C_i and C_j belong to G_i or G_j, D_{ii} will be full rank. However if $C_i \in G_i$ and $C_j \in G_{j \ne i}$, D_{ij} loses rank property. To restore full-diversity, schemes based on rotations of information symbols have been proposed (see e.g. [3], [7]). In this paper, we propose to rotate constellation and scale amplitude transmitted signals in s' by a suitable angle θ and λ to guarantee full-diversity for maximizing the CGD. For a unit-radius QPSK constellation, the rotating result is the same with signal constellation for transmission. Consider two different codewords $C_i, C_j \in C$. In order to ensure full spatial diversity, the difference matrix $C_i - C_j = D_{ij}$ between any two distinct codewords in the extend set S must be full rank [2]. When both codewords C_i and C_j belong to the G_0 or G_1, D_{ij} will be full rank. However, if $C_i \in G_0$ and $C_j \in G_1$ (or vice versa), D_{ij} lose full rank property. To restore full-diversity, the scheme based on rotations of information symbols has been proposed [4]. For a unit-radius QPSK constellation, the different angle should guarantee that the proposed constellation is different from the traditional constellation. Then, the coding gain matrices between two different code works should be full rank to achieve the higher determinant as large as possible. The optimum combination of angle θ and λ should follow such rule to design the constellations.

B. Optimum Values of Angle θ and Amplitude λ

With different angle θ and λ values, the system performance should be different. The main idea to introducing the angle θ and λ is to ensure full diversity for the proposed high-rate STBC. One important selection criteria for θ and λ is to maximize the determinate of the CG with different combinations of θ and λ. In addition, CG is defined as the minimum product of nonzero singular values of D_{ij} for overall distinct codewords pairs. In order to ensure full spatial diversity, the codeword difference matrix

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

$\det\left(\left(C_i - C_j\right)^* \left(C_i - C_j\right)\right)$ between any two distinct codewords in the extended set S must be full rank.

$$\det\left(\left(C_i - C_j\right)^* \left(C_i - C_j\right)\right) \hat{=}$$

$$\det\left(\begin{pmatrix} 1-\lambda e^{j\theta} & 1-\lambda e^{j\theta} \\ 1-\lambda e^{-j\theta} & 1-\lambda e^{-j\theta} \end{pmatrix}^* \begin{pmatrix} 1-\lambda e^{j\theta} & 1-\lambda e^{j\theta} \\ 1-\lambda e^{-j\theta} & 1-\lambda e^{-j\theta} \end{pmatrix}\right) \quad (4)$$

While difference matrix $C_i - C_j = D_{ij}$ is a full rank matrix, $\det(D_{ij} D_{ij}^*) \neq 0$. The optimum value of angle θ and λ while $\det(D_{ij} D_{ij}^*)$ is maximized by using the first derivative property.

$$\begin{cases} \dfrac{\mathrm{d}\left(\det(\mathrm{D_{ij}D_{ij}^*})\right)}{\mathrm{d}\theta} = 0 \\ \dfrac{\mathrm{d}\left(\det(\mathrm{D_{ij}D_{ij}^*})\right)}{\mathrm{d}\lambda} = 0 \end{cases} \Rightarrow \begin{cases} \lambda(\sin\theta - \cos\theta) = 0 \\ \lambda^2 - \lambda\sin\theta = 0 \end{cases}, \begin{cases} \theta = \pi/4 \\ \lambda = 1/\sqrt{2} \end{cases} \quad (5)$$

Here, the optimum value of angle θ and λ are derived from these formulas. The candidate of QPSK signal constellation can be obtained by the optimum values of angle θ and λ. Then, the traditional QPSK signal constellation is the collection $s \in \left\{1/\sqrt{2} \cdot (\pm 1 \pm i)\right\}$, and the proposed constellation also exits in another collection $s' \in \lambda s \exp(j\theta) = \left\{\pm 1/2, \pm i\right\}$.

C. Low Complexity Decoding Algorithm

The output symbols $R = \left[r_1, r_2\right]^T$ received vector over two consecutive symbol periods can be represented as follows:

$$\mathbf{R} = HS_{G_i} + N \ , \ \mathbf{H} = \begin{bmatrix} h_1 & h_2 \\ -h_2^* & h_1^* \end{bmatrix} \quad (6)$$

where $\mathbf{R}$ are 1×2 complex matrix representations of the received signals corresponding to the transmitted codewords in the form of G_i and $G_{j \neq i}$, respectively. The path gains from the two transmit antennas to the mobile is $\mathbf{H}$ that are independent and identically distributed (i.i.d) with $h_i \sim CN(0, \sigma^2)$ and $E(h_i h_j) = 0$ if $i \neq j$. The channel is assumed to be known perfectly at the receiver and the noise samples $N = [N_1, N_2]^T$ are independent samples of a zero-mean complex Gaussian random variable. The channel matrix $\mathbf{H}$ is a quaternion and we have $HH^* = \left(|h_1|^2 + |h_2|^2\right) I_2$. Two simple matched-filtering operations, $H^* G_i$ and $H^* G_{i \neq j}$ are performed to generate two candidate solutions, namely, $\hat{S}_i$ and $\hat{S}_{j \neq i}$ which are then compared using the metric

$$\tilde{S} = \arg\min_{s} \left\| \left[r_1, r_2\right]^T - HS \right\|^2 \quad (7)$$

The decoding result for $b0$ follows directly once the decision between S_i or $S_{j \neq i}$ is made. Now we begin to state the decoding algorithm for the proposed high-rate scheme. At first, the transmitted codewords can be divided into two parts of S, which can be expressed as mathematical expression $S = \left\{S_1, S_2\right\}$. These two different subsets S_1, S_2 within the set can be used to transmit $G_{1,i} \in S_1, \forall i$ and $G_{2,i} \in S_2, \forall i$ with the additional bit for transmission, respectively. The decoding algorithm can be divided into three steps by using (7)

1. Following the decoding criterion formula (7), one constellation of set S_1, S_2 is selected to decode the transmitted codes through the received vector. Without loss of generality, at the first step, S_1 is selected to decode and note the metric of the decoding result and $d_1 = \left\| \left[r_1, r_2\right]^T - H\tilde{G}_1 \right\|^2$, $\tilde{G}_1$ is the decoding result in our decoding procedure.
$$\tilde{G}_1 = \arg\min_{S_1} \left\| \left[r_1, r_2\right]^T - HS_1 \right\|^2$$
2. At this step, another subset S_2 is selected to decode the transmitted codes as the same procedure with step 1, and note the metric of this step $d_2 = \left\| \left[r_1, r_2\right]^T - H\tilde{G}_2 \right\|^2$
3. To compare with the metrics of previous two steps and select the maximum value between two metrics, the transmitted codes and the additional bit can be obtained through the three steps.

Comments: Although the proposed algorithm can improve the system transmission rate, the decoding algorithm should be more complicated then the original algorithm with more adders and multipliers achieving high rate property.

IV. EXTENTION TO 4 ANTENNAS CASE

Consider the following example of a candidate of the STBC with rate-1 and full-diversity complex quasi-orthogonal design based on Quasi-orthogonal STBC proposed by Jafarkhani [5]. In our proposition, we extend the set of transmitted signal constellation by a multiplied factor of $\lambda \exp(j\theta)$ to the quasi-orthogonal entrance selected from the conventional signal constellation.

$$G = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ -x_2^* & x_1^* & -x_4^* & x_3^* \\ -x_3^* & -x_4^* & x_1^* & x_2^* \\ x_4 & -x_3 & -x_2 & x_1 \end{bmatrix} \quad (8)$$

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

We expand the signaling set and increase the rate to $9/8$ (for QPSK) by considering the following signal multiplied by the multiplying the factor $\lambda \exp(j\theta)$.

$$G_0(x_1,x_2,x_3,x_4,b_0=0)=\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ -x_2^* & x_1^* & -x_4^* & x_3^* \\ -x_3^* & -x_4^* & x_1^* & x_2^* \\ x_4 & -x_3 & -x_2 & x_1 \end{bmatrix} \quad (9)$$

$$G_1(x_1,x_2,x_3,x_4,b_0=1)=\begin{bmatrix} x_1e^{j\theta} & x_2e^{j\theta} & x_3e^{j\theta} & x_4e^{j\theta} \\ -x_2^*e^{-j\theta} & x_1^*e^{-j\theta} & -x_4^*e^{-j\theta} & x_3^*e^{-j\theta} \\ -x_3^*e^{-j\theta} & -x_4^*e^{-j\theta} & x_1^*e^{-j\theta} & x_2^*e^{-j\theta} \\ x_4e^{j\theta} & -x_3e^{j\theta} & -x_2e^{j\theta} & x_1e^{j\theta} \end{bmatrix}$$

To compare these different STBCs, the transmission scheme is almost the same with the two antennas case with different rotating angles about the original signal constellation.

The CGD (coding gain distance) between a pair of codewords C=G(s_1,s_2,s_3,s_4) and C'=G(s'_1,s'_2,s'_3,s'_4) from the QOSTBC is given by

$$\begin{aligned} &\mathrm{CGD}(C,C')=\det\left[\mathrm{D}(C,C')^{H}\mathrm{D}(C,C')\right] \\ &=\left(\left|(s_1-s'_1)-(s_4-s'_4)\right|^2+\left|(s_2-s'_2)+(s_3-s'_3)\right|^2\right)^2 \\ &\cdot\left(\left|(s_1-s'_1)+(s_4-s'_4)\right|^2+\left|(s_2-s'_2)-(s_3-s'_3)\right|^2\right)^2 \end{aligned} \quad (10)$$

Then, in order to simplify these formulas, we replace x_1 with s_1-s'_1, x_2 with s_2-s'_2, x_3 with s_3-s'_3, and x_4 with s_4-s'_4

$$\det(G^HG)=\left(|x_1-x_4|^2+|x_2+x_3|^2\right)^2\left(|x_1+x_4|^2+|x_2-x_3|^2\right)^2 \quad (11)$$

To derive the optimum values of angle θ and λ, the simpler first derivative of $\det(G^HG)$ also calculates.

$$\begin{cases} \dfrac{d(\det(G^HG))}{d\lambda}=\lambda^3(\cos^4\theta-\sin^2\theta\cos^2\theta) \\ \dfrac{d(\det(G^HG))}{d\theta}=3\lambda^4\cos^3\theta\sin\theta-\lambda^2\sin\cos^2\theta \end{cases} \quad (12)$$

$$\Rightarrow \theta=\pi,\lambda=1/\sqrt{3}$$

Here, a similar way can be used to calculate the optimum values of angle θ and λ, the values derived are used to construct the new transmitted signal constellation according our proposed motivation in this paper.

V. NUMERICAL RESULTS

In this section, we show some simulation results for the proposed transmission scheme. Fig. 2 shows the BER performances versus SNR to compare the proposed design with the Almouti code and Quasi-orthogonal [5] equipped with 2 or 4 transmitting and receiving antennas. The proposed transmission scheme simulated in this figure has a similar performance with Quasi-orthogonal above 20dB. QPSK modulation, 8PSK modulation and flat fading channel are considered in all simulations. In the high SNR region 20-25 dB of Fig. 1, the BER performance of proposed scheme with a high rate-9/8 transmission is similar to Quasi-orthogonal code, but there also exists a visible gap in the low SNR region 10-20dB. Moreover, the BER performance of proposed design is almost 1-2dB away from the conventional Almouti code and Quasi-orthogonal at the BER of 10^{-5}.

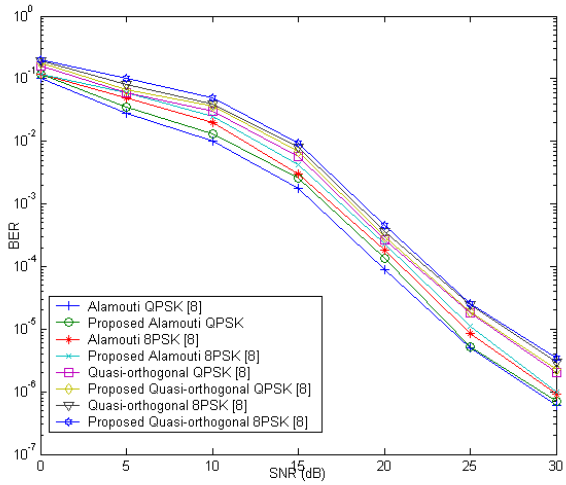


Fig. 2 Comparisons of FER Performance

In Fig. 3, we compare our proposed rate- 5/4 code with the Alamouti [1] code and rate- 9/8 code with the Quasi-orthogonal code [6] using the measure of Effective Throughput η defined as:

$$\eta=(1-FER)*R*\log_2(M), \quad (13)$$

where R is the code rate, and M is the constellation size, and FER denotes the frame error rate, and means the frame error rate, and each frame contains 4 symbols (9 bits plus one additional bit). The Fig. 3 shows that at high SNR, our code achieves a higher throughput level of channel use (PCU) whereas the achievable throughput for the Alamouti code [1] PCU. Fig. 3 also depicts the achievable throughput of a $rate-9/8$ code that uses pure rotations to ensure full-diversity and maximize coding gain. The proposed transmission scheme achieves 2.25 bits per channel transmission whereas the effective throughput of the QOSTBC is 2 bits. A crossing point exits at a SNR level of 20dB. Similarly, the effective throughput performance of high rate-9/8 [8] is also simulated in this figure. From 20dB to 30dB of SNR, the effective throughput has been improved compared to [8] due to low FER in the proposed design. This result is matched with the conclusion of Fig. 1 in which the system performance is affected by the CGD of coding gain matrix.

It demonstrates that optimum rotation achieves higher performance for all values of SNR. The selective angle rotation ensures full-diversity and high rate at the cost of reduced coding gain. It is also possible to turn this coding loss into gain by introducing block codes for the fading

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

channel as in [7]. For an overall rate of 1, this combination of coding techniques outperforms the Alamouti code [1] at the price of higher decoding complexity.

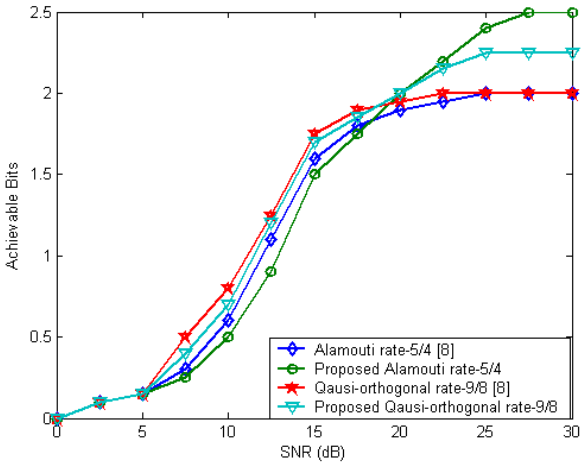


Fig. 3 Comparisons of Throughput Performance

VI. CONCLUSIONS

We exploited the algebraic structure of quaternions to design and optimize a novel high-rate, full-diversity STBC for 2 and 4 transmitting antennas. We introduced the concept of selective power scaling to guarantee full diversity for the designed code. The angle rotation was further optimized to maximize available coding gain. The computer simulations show that the effective throughput is improved in the high SNR region.

ACKNOWLEDGMENT

This work was supported by World Class University R32-2009-000-2014-0 NRF, and Fundamental Research (FR) 2010-0020942 NRF, Korea, and PH.D program and New seedling foundations of Guangzhou University, and Natural Science Foundation of Guangdong Province (S2011040004068), China.

REFERENCES

[1] S. M. Alamouti, “A simple transmit diversity technique for wireless communications,” IEEE J. Select. Areas Commun., vol. 16, pp. 1451–1458, Oct. 1998.

[2] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, “Space-time block codes from orthogonal designs,” IEEE Trans. Inform. Theory, vol. 45, pp. 1456–1467, July 1999.

[3] B. Hassibi, B. M. Hochwald, A. Shokrollahi, and W. Sweldens, Representation theory for high-rate multiple-antenna code design, IEEE Trans. Inform. Theory, vol. 47, pp. 2335-2367, Sept. 2001.

[4] H. Jafarkhani and N. Seshadri, “Super-orthogonal space-time trellis codes,” IEEE Trans. Info. Theory, vol. 49, pp. 937–950, April 2003.

[5] H. Jafarkhani, “A quasi-rothogonal space time block code,” IEEE Trans. Commun., vo. 49, pp. 1-4, Jan. 2001.

[6] W. Su and X.-G. Xia, “Signal constellations for quasi-orthogonal spacetime block codes with full diversity,” IEEE Trans. Inform. Theory, vol. 50, pp. 2331–2347, Oct. 2004

[7] M. Z. A. Khan and B. S. Rajan, “Single-symbol maximum likelihood decodable linear STBCs,” IEEE Trans. Info. Theory, vol. 52, pp. 2062–2092, May 2006.

[8] S.Das, N.Al-Dhahire, and R.Calderbank, “Novel Full-Diversity High-Rate STBC For 2 and 4 Transmit Antennas,” IEEE Trans.Commun.vol.10, No.3, March 2006

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Complexity and Fairness Analysis of a new Scheduling Scheme for VoIP in 3G LTE

Richard Musabe, Hadi Larijani
Glasgow Caledonian University
Glasgow, Scotland, UK
e-mail: richard.musabe@gcu.ac.uk, h.larijani@gcu.ac.uk

Abstract— **3G Long Term Evolution is an emerging and promising technology that aims at providing broadband ubiquitous internet access and improving multimedia services. This is achieved through streamlining the system for packet services since long term evolution is an all Internet protocol based network. The fact that 3G long term evolution is a packet based network brings along some improvements in the form of higher bit rates, lower latencies, and a variety of service offerings. However, some technical challenges are expected to arise when voice traffic is transmitted over a long term evolution network. This has become an interesting area of research and different types of resource management schemes have been developed which are quite challenging and complex. In this paper, we analyze the complexity and fairness of our proposed scheduling scheme for voice over internet protocol in 3G long term evolution called voice over internet protocol optimization scheduling algorithm. We compare it with other algorithms in literature. There is second order complexity in the number of users based on quality feedback, queue length metrics, and there is linear complexity in resource blocks using voice over internet protocol optimization scheduling algorithm. Simulation results also showed approximately 10 – 20 percent improvement in fairness and performance based on the fairness index and throughput.**

Keywords-LTE; Scheduling Schemes; VoIP; Complexity; Fairness.

I. INTRODUCTION

3G Long Term Evolution (LTE) was identified by the third generation partnership project (3GPP) as the preliminary version of next generation wireless communication systems because of its high data rates [1]. This mobile cellular communications technology provides a maximum 100Mbps downlink and 50Mbps uplink when using 20 MHz bandwidth [2]. In the downlink physical layer, LTE uses Orthogonal Frequency-Division Multiple Access (OFDMA) radio technology to meet the LTE requirements for spectrum flexibility and enables cost-efficient solutions for wide carriers with high peak rates. In the uplink, LTE uses a pre-coded version of OFDMA which is Single-Carrier Frequency-Division Multiple Access (SCFDMA) in order to compensate for a drawback with normal OFDMA which has a high Peak-to-Average-Power Ratio (PAPR) [3].

Wireless technology has expanded from voice only to high-speed data, multimedia applications, and wireless internet [4]. LTE requirements for high data rates are achieved by the fact that this technology is only designed for packet switched networks (PSN); hence, there is no need for the circuit switched mode. However, this design brings with it more technical challenges especially for voice services. Voice over internet protocol **(**VoIP) services are both delay and packet loss sensitive. The biggest challenge of VoIP over LTE is the delivered Quality of Service (QoS). Normally, users would expect voice with the same quality as that provided by circuit switched networks. However, traffic delivered over PSNs is subject to delay and packet loss [5]. A major issue with VoIP over LTE is that 3G LTE adopts a different method of resource transmission from other cellular systems like Code Division Multiple Access (CDMA). It uses Physical Resource Blocks (PRB) as its transmission unit. PRBs can be defined as the basic unit with both frequency and time aspects [6]. Basically, the base station of 3G LTE, known as eNodeB has a fixed number of available PRBs according to their allocated bandwidth and it is supposed to assign PRBs repeatedly at every Transmission Time Interval (TTI) [2].

Different techniques have been introduced in the literature to over come the challenges faced when real time traffic is trasmitted over an LTE network. In [7], Yaacoub, Al-Asadi, and Dawy proposed two low complexity heuristic algorithms. The complexity of both algorithms was analysed. The first algorithm had a linear complexity in the number of users and a quandratic complexity in the number of resource blocks. The second algorithm had a linear complexity in both the number of user and resource blocks. It was shown that good results could be achieved by the proposed linear complexity algorithm (second algorithm). It was also shown through simulations that the maximization of sum throughput leads to a higher cell throughput, although considering the logarithm of throughput as a utility function ensures proportional fairness, and thus constitutes a tradeoff between throughput and fairness.

In [8], Zhao et al. investigated two fairness criterias with regards to adaptive resource allocation for uplink OFDMA systems. These two criteria were Nash bargaining solution (NBS) fairness and proportional fairness (PF). These two

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

criterias can provide attractive tradeoffs between total throughput and each user's capacity. Using Karush-Kuhn-Tucker (KKT) condition and iterative method, two effective algorithms were designed to achieve NBS fairness and proportional fairness respectively. Through simulation results, NBS fairness criteria shows better performance in total capacity but the BS cannot control the rate ratio because it only depends on the channel state of the users. PF Criteria can provide a controllable rate ratio regardless of the channel condition for each user. However, to achieve the hard fairness, the system capacity degrades sharply.

In [9], Piro et al. proposed a new open-source framework to simulate LTE networks. In this simulator, different scheduling algorithms were developed, these include; proportional fair (PF), exponential proportional fair(EXP-PF), and modified largest weighted delay first (M-LWDF). We will consider the first two algorithms since their fairness and complexity context constitues an extension to algorithms described in [7][8]. We will also compare the performance of these two algorithms to our proposed scheduling algorithm VOSA. we will refer to these two algorithms in the simulations as PF and EXP-PF. Our involvement in [7] is that we only compared the complexity and fairness of our algorithm to algorithm 1 in [7]. Since algorithms in [7] are extension to the algorithms proposed in [9], we also compared the performance of our algorithm to those proposed in [9].

PF: This scheduler was developed in [9] and its main aim is to maximize the total network throughput and to guarantee fairness among flows. It assigns radio resources taking into account both the experienced channel quality and the past user throughput [10]. This scheduler uses the metric which is defined as the ratio between the instantaneous available data rate and the average past rate with reference to the i-th flow in the j-th flow subchannel. This can be depicted in equation 1 obtained from [9].

$$W_{i,j} = \frac{r_{i,j}}{R_{i,j}} \quad (1)$$

where $W_{i,j}$ is the scheduler metric, $R_{i,j}$ is the estimated average data rate, and $r_{i,j}$ is the instantaneous available data rate which is computed by the AMC module, considering the channel quality indicator (CQI) feedback that the UE hosting the i-th flow have sent for the j-th subchannel. It should be also noted that i and j are sub channel flows.

EXP-PF: This scheduler was also developed in [9] and it basically aimed at increasing the priority of real-time flows with respect to non-real-time flows, where their head-of-line packet delay is very close to the delay threshold [11]. Its metrics were computed in [9] using the following equations.

$$W_{i,j} = \exp\left(\frac{\alpha_i D_{HOL,i} - X}{1+\sqrt{X}}\right)\frac{r_{i,j}}{R_{i,j}} \quad (2)$$

and

$$X = \frac{1}{N_{r,t}} \sum_{i=1}^{N_{r,t}} \alpha_i D_{HOL,i} \quad (3)$$

with N_{rt} being the number of active downlink real-time flow. Considering a packet delay threshold T_i, the probability σ_i is defined as the maximum probability that the delay $D_{HOL,I}$ of the head-of-line packet delay exceeds the delay threshold.

Therefore, α_i is given by;

$$\alpha_i = -\frac{\log \alpha_i}{T_i} \quad (4)$$

Equations 3 and 4 proposed in [9], calculates the average summation of the entire down link real time flows based on the probability that the first packet to be transmitted in the queue exceeds the delay threshold. This helps to prioritize down link real time flows.

With all these techniques introduced in the literature, there are still some challenges when real-time traffic like voice is transmitted over an LTE network. This is mostly due to the fading channels of wireless links and the delay and packet loss sensitive voice characteristic. Another issue is that, most of the proposed solutions in the literature are found to be more complex and do not grant fairness between VoIP users. Users which are very close to the base station are assigned more PRBs than those far from the base station.

So, in this work, we analyze the complexity and fairness of our proposed scheduling scheme for VoIP in 3G LTE called VoIP Optimization Scheduling Algorithm (VOSA) [12]. Then compare it with other scheduling algorithms in [8], which are PF and EXP-PF, in order to analyze its performance based on user throughput. The main contribution in this paper that was not discussed in our previous publication is that we analyzed the complexity, fairness, and throughput of our model presented in [12].

The simulation results were generated using the open source LTE system simulator called LTE-SIM [9]. It models different uplink and downlink scheduling strategies in multicell/multiuser environments; taking into account user mobility, radio resource optimization, frequency reuse techniques, the adaptive modulation, and coding (AMC) module. It also includes other aspects that are relevant to the industrial and scientific communities.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Our contributions in this paper are:

- Complexity and fairness analysis of our proposed scheduling algorithm VOSA and compared it with algorithm 1 in [7].
- Performance analysis of VOSA and compared it with algorithms in [9].

The rest of the paper is organised as follows: Section II discusses the VOSA Scheduling algorithm, metric maximization, and describes in summary VOSA algorithm. Section III describes the simulation. Section IV presents complexity and fairness analysis, as well as performance simulation results. Section V reviews the main conclusions.

II. VOSA SCHEDULING ALGORITHM

Our proposed scheduling algorithm and its details can be found in [12]. The main aim of this proposed scheduling algorithm is to improve the QoS of voice traffic when transmitted over an LTE network. At the same time it reduces the negative impact, which may be caused by the introduction of the new algorithm on the entire system's performance. This algorithm is activated at every TTI by considering if there is a VoIP call and if the duration period of the new algorithm has not exceeded the limit. To determine the duration of our new algorithm, we use the adaptive method proposed in [2]. This method provides limits to the *VOSA* which is adaptively changed between a pre-specific minimum and maximum value according to the ratio of dropped packets. Higher drop ratio means that there are many ongoing VoIP calls, and hence, it is necessary to increase the limits to allow more consecutive TTIs to be dedicated to VoIP calls. On the other hand, a low drop ratio implies that QoS of VoIP calls are satisfied at decent levels, and thus, it is safe to reduce the duration of the algorithm and serve other service in the network.

Our scheduling scheme is designed by making modification to the algorithm in [2]. *VOSA* allocates PRBs to VoIP calls based on the arrival time metric. Once the PRBs allocation is done, the scheduling order of the calls is determined by the size of the following factors: Quality feedback (QF) and Queue length (QL) of each call. The better the factor values are, the earlier the corresponding call is scheduled.

A. Metric Maximization

Let $N_{RB,K}$ be the number of resource blocks allocated to the number of users K, T_k be the arrival time associated to k user and $QL_{(k)}$ be the length of k's queue. Every user k sends back the quality feedback value $QF_{(k)} \in \{1,\ldots\ldots, QF^{(max)}\}^{k*1}$ containing supported values for the user k. The maximization of user utility metrics can be formulated as follows:

$$MAX \sum_{n=1}^{K} U\left(\frac{QF_k QL_k}{N_{RB,k}}\right) \quad (5)$$

where $QL_{(k)} \geq 1$

U ($QF_{(k)}$ $QL_{(k)}$) is the user utility as a function of two main metrics (QF and QL), given the allocation of resource block $H_{RB,N}$ to user k.

1. Quality Feedback:

In order to obtain quality feedback metric, we used the Time-domain proportional fair method [TD-PF] [13] and it is obtained from the equation below.

$$QF_{k,j}[t] = \frac{R_{k,j}[t]}{Th_{k,j}[t]} \quad (6)$$

where:

$QF_{k,j}[t]$ - Quality feedback Metric for user 'k' in the channel 'j' in the instant't'

$R_{k,j}[t]$ - Shannon Channel Quality Indicator (CQI) of user 'k' in the channel 'j' in the instant 't'

$Th_k[t]$ - Average delivered user throughput, it is calculated based on the transmitted signal's SINR

2. Queue Length**:**

In order to obtain Queue length metric, we adopted the queuing method in the LTE-SIM simulator. Different traffic generators were developed, these generated packets that are transported by a dedicated radio bearer at the application layer. Using the application class, we were able to generate the packets and deliver them to the network. Once the packets reach the network, they are forwarded to the user-plane protocol stack to add protocol headers.

Then, the packets are placed in the queue by the MAC queue class at the MAC layer before being sent to the destination. The MAC queue object have got a counter which increases or decreases when the packet is inserted or removed from the queue respectively. Based on the counter in the MAC queue object, the queue length metric is determined. It should be noted that different MAC queue objects can be created in order to facilitate different traffic types.

QL_k = m_queuesize+N_datapackets*8 (7)

where 8 is the packet overhead due to, Radio Link Control (RLC) (2bytes), MAC headers (3 bytes) and Cyclic Redundancy Check (CRC) (3 bytes).

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

B. *Summary of VOSA Algorithm*

This algorithm performs the scheduling operation based on the user utility metrics and the better the metrics are, the earlier the call is scheduled. Its operation includes the following steps:

- Indentify the traffic type whether voice or any other traffic
- Determine the user utility metrics (QF,QL)
- Find the user with the highest user utility metric as defined in equation 5
- Consider the set of available resource blocks RBs N_{avail_RB}, at every start of the algorithm , N_{avail_RB} = {1,2,……., N_{RB}}
- Assign the resource block N* to the user k* with the highest user utility metrics value such that $N_{RB,}k^*$ = $N_{RB,}k^*$ U {N*}
- Schedule the user k* first
- Delete the user k* and resource block N* from their respective lists
- Repeat all the steps until all users are scheduled and if more resource block exists then allocate them to other traffic types

III. SIMULATION SETUP

A. *PRB Characteristics*

In this sub-section, we introduce the characteristics of PRBs, which are the transmission resources. LTE systems consists of both a time and a frequecy planes. The time plane is divided into 1 ms TTI, which consists of two slots of 0.5 ms to form 1 ms sub frames, where each sub frame contains 7 OFDMA symbols.

In each TTI, there are 14 OFDMA symbols, where 2 symbols out of 14 are reserved for uplink pilot transmission, while the other 12 symbols are used for data and control information transimission. TTI can be defined as the minimum allocation unit in the time domain [12]. If we consider the frequency plane, the minimum allocation unit is the PRB, where each PRB contains 12 subcarriers of 15 Khz bandwidth each.

The number of OFDMA symbols in a resource block depends on a cyclic prefix being used. All these can be depicted in Fig. 1. It must be noted that VoIP packets must be transmitted per TTI and they can occupy one or more PRBs [5]. The amount of data bits that can be transmitted by one PRB depends on the link between the eNodeB and the user mobile terminal. This is due to the fact that 3G LTE uses adaptive modulation and coding (AMC), which changes modulation and coding schemes depending on the wireless link conditions.

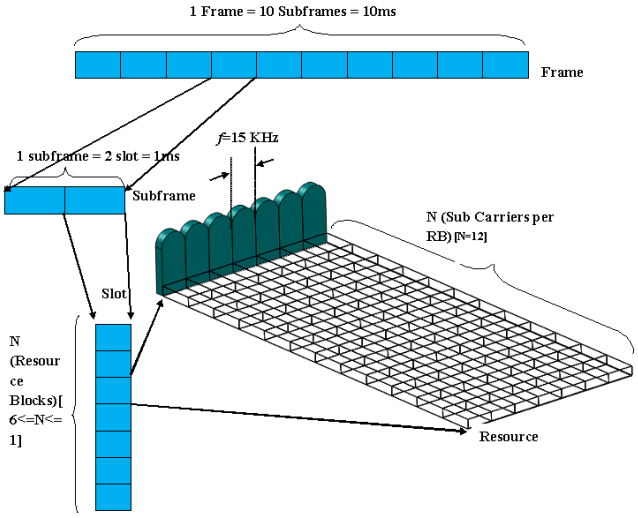


Figure 1. The structure and allocation of the eNodeB transmission resources symbols.

B. *Scenario Setup*

Our network topology is made up of a set of cells and different network nodes which include; the EnodeB, mobility management/gateway (MME/GW), and user equipements (UEs). All the simulations were run in a three tier diamond-pattern macro scenario with 19-3-sector sites which totaled to about 57 cells. Most of the simulation parameters are presented in the table 1 below. VoIP flows are generated by the traffic generator in LTE-SIM called VoIP application which generates G.729 voice flows. The voice flow has been modelled with an ON/OFF Markov chain. The ON period is exponentially distributed with a mean value of 3s and the OFF period has a truncated exponential probability density function with an upper limit of 6.9s as well as an average value of 3s [9].

During the ON period, the source sends 20bytes sized packets every 20 ms which implies that the source data rate is 8 kb/s, on the other hand during the OFF period the rate is zero because the presence of voice activity detector is assumed. Three different scheduling algorithms were used in all simulation scenarios, these are: our proposed VOSA as well as EXP-PF and PF developed in [9].

IV. DISCUSSION AND PERFORMANCE ANALYSIS

A. *Complexity Analysis*

Our proposed algorithm performs the scheduling operation after searching the user with highest utility metrics based on QF and QL. Therefore, the complexity to schedule the first user is O(KN), this will be the complexity for the first iteration. The complexity to schedule the second user is O((K-1)N) and so on.

In our algorithm, the number of iterations depends on the number of users K. As there are K iterations, the overall algorithm complexity can approximately be expressed as

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

$O(K^2N)$. This implies that there is a second order complexity in the number of users based on QF, QL metrics and there is also linear complexity in resource blocks N. This is due to the fact that there is no search done on the resource blocks, any available resource block is assigned to the user with highest metric.

If we compare our algorithm to algorithm 1 in [7] that has a linear complexity in the number of user and quadratic complexity in the number of resource block, i.e., '$O(N^2K)$', it is clear that our algorithm will only outperform it when the number of users are few since it will perform less iterations however when the number of users increases, algorithm 1 in [7] performs better.

TABLE 1. SIMULATION PARAMETERS

Simulation Parameters	Values
Bandwidth	5MHZ
PRB Structure	12subcarriers,2subframes
TTI	1msec
Number of available PRBs	25
Modulations for AMC	QPSK
Number of sectors	3
Simulation time	1000 TTIs
Cyclic prefix	Normal
Scheduling algorithms	VOSA,EXP-PF, and PF
Cell radius	1 km

B. *Fairness Analysis*

The fairness aspect is introduced mainly to solve the resource starvation problem, where users close to the base station are allocated more resources and edge users generally suffer from resource starvation [7]. Fairness can be describes as a loose concept which implies that all users are allocated equal amount of resources in order to meet the QoS requirements. From the fairness point of view, we compared our algorithm with PF and EXP-PF developed in [9]. Their fairness and complexity context constitues an extension to algorithms described in [7][8] and they are also the bench mark schedulers in the simulator that we used. We measured the fairness index of all the scheduling schemes. As seen in Fig. 2, fairness index decreases as the number of users increases. The fairness index of VOSA is higher than that of PF but lower than EXP-PF.

It should be noted that the main advantage of VOSA scheduling algorithm is to improve the QoS of voice traffic when transmitted over an LTE network. At the same time it reduces the negative impact which may be caused by the introduction of the new algorithm on the entire system's performance. However, when we consider fairness and performance analysis, EXP-PF out performs VOSA due to the following reasons;

EXP-PF employs the fairness concept in [7], which uses the algorithmic utility function that is associated with proportional fairness of the utility based optimization. This helps in achieving a better fairness factor. Also, during the allocation scheme, EXP-PF erases all the packets belonging to the real time flow from the MAC queue if they are not transmitted before their deadline expiration. This helps to improve its performance by avoiding the waste of resources such as bandwidth.

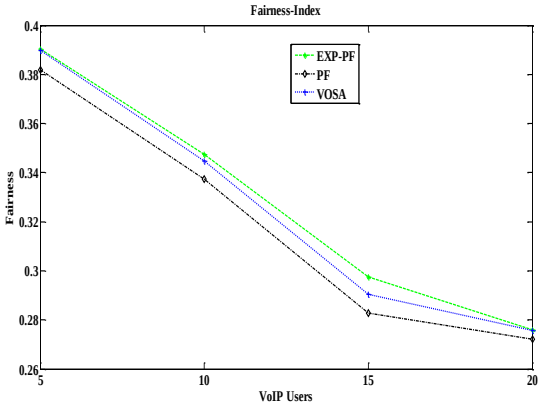


Figure 2. Fairness Index Comparison

C. *Performance Analysis*

Performance analysis was made by measuring user throughput for all three schedulers. As it can be seen in Fig.3, throughput decreased as the number of VoIP users increased in all algorithms.This is mainly due to the fact that some VoIP packets were being dropped as the number of users were being increased which resulted in the less utilisation of all assigned PRBs.

It is well known that VoIP packets are small packets; hence, many packets are needed to fully utilize the available PRBs. However, as congestion increased in the network, it led to VoIP packets to be dropped which led to less utilization of the available PRBs.

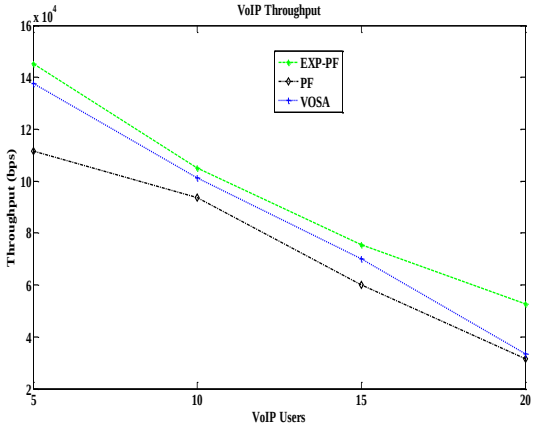


Figure 3. Throughput Comparison

V. CONCLUSION AND FUTURE WORK

In this paper, we analyzed the complexity and fairness factor of our proposed scheduling algorithm VOSA. Through simulations, we were able to compare it with other scheduling algorithms in literature. It was analyzed that VOSA performs better when the number of users is small since it schedules users after searching the user with highest utility metrics based on QF and QL and the search goes on for all available users. So the small the number of users, the few search iterations done and hence the better performance. However, VOSA performed better than PF but slightly lower than EXP-PF.

In future work, we will try to employ different tests such as real life scenarios in order to analyze the practicability of our results and to make them more reasonable. We are also working on the way of optimizing our scheduling algorithm such that we lower its complexity by just performing one main search throughout all the available users and store the data for each user separately. This would make it more scalable since a single iteration is performed.

REFERENCES

[1] S. Y. Kim, "An Efficient Scheduling Scheme to Enhance the Capacity of VoIP Services in Evolved UTRA Uplink," EURASIP journal of Wireless Communications and Networking, vol 2008, Mar. 2008, pp.1-9, doi:10.1155/2008/732418.

[2] S. Choi, K. Jun, Y. Shin, S. Kang, and B. Choi, "Mac Scheduling Scheme for VoIP Traffic Services in 3G LTE," IEEE 66th Vehicular Technology Conference, pp. 1441-1445, Oct. 2007, doi: 10.1109/VETECF.2007.307.

[3] Rohde and Schwarz, "*UMTS* long term evolution (LTE) Technology introduction," A report by Rohde and Schwarz, pp. 1-30, Mar. 2007.

[4] M. C. Chuah, and Q. Zhang, Introduction to Wireless Communications, US: Springer, 2006.

[5] S. Saha and R. Quazi, "Priority-Coupling- A Semi-Persistent MAC Scheduling Scheme for VoIP Traffic on 3G LTE," ConTEL,10th International Conference on Telecommunications, pp. 325-329, Aug. 2009.

[6] 3GPP, Physical layer aspects for Evolved UTRA, 3GPP Technical report 25.814, version7.1.0, pp. 1-135, Sep. 2006.

[7] E.Yaacoub, H. Al-Asadi, and Z. Dawy, "Low Complexity Scheduling Algorithms for LTE Uplink," Computers and Communications, ISCC 2009, IEEE Symposium, pp. 266 - 270, July 2009 doi: 10.1109/ISCC.2009.5202296.

[8] Y. Zhao, L. K. Zeng, G.Xie, Y.A. Liu, and F. Xiong," Fairness based resource allocation for uplink OFDMA systems," Journal of China universities of post and telecommunications, vol. 15, No. 2, June 2008, pp. 50 – 55.

[9] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE Cellular Systems: An Open-Source Framework," IEEE Transactions on Vehicular Technology, vol. 60, no. 2, pp. 1-16, Feb. 2011.

[10] G. J. Choi and S. Bahk, "Cell-throughput Analysis of the proportional fair scheduler in the single-cell environment," IEEE Transaction Vehicular Technology, vol. 56, no. 2, April 2007, pp. 766-778, doi: 10.1109/TVT.2006.889570.

[11] R. Basukala, H. M. Ramli, and K. Sandrasegaran, "Performance analysis of EXP/PF and M-LWDF in downlink 3GPP LTE systems," 1st AH-ICI . Kathmandu, Nepal, Nov. 2009, pp. 1-5, doi: 10.1109/AHICI.2009.5340336.

[12] R. Musabe, H. Larijani, B. Stewart, and T. Boutaleb, "A New Scheduling Scheme for Voice Awareness in 3G LTE," IEEE Computer Society, Sixth International Conference on Broadband and Wireless Computing, Communication and Applications, Dec. 2011, pp. 1-8, doi: 10.1109/BWCCA.2011.46.

[13] J. A. Rodriguez, "Radio Resource Management Centralized for Relayed Enhanced LTE-Networks," A report by the department of electronics and information systems, Aalborg University, pp. 1 – 40, June, 2009.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Optimal Bandwidth Consumption for IPTV Services over WiMAX Multihop Relay Networks

Mohamed-el-Amine Brahmia, Abdelhafid Abouaissa and Pascal Lorenz

University of Haute Alsace – MIPS-GRTC, 34, Rue de Grillenbreit 68000 Colmar – France

{mohamed-el-amine.brahmia, abdelhafid.abouaissa, pascal.lorenz}@uha.fr

***Abstract*—The quick evolution in technologies has allowed IPTV video stream delivery over IP networks. For that reason consumers have anticipated predictions in which the evolution of IP-based next-generation networks may be eventually driven by video service delivery requirements. An IEEE 802.16j mobile WiMAX relay network is a next-generation mobile wireless broadband network. Compared to IEEE 802.16e which also supports mobility, IEEE 802.16j introduces relay stations to the network to offer improved coverage and capacity over multihop radio systems. However, to supply different IPTV services (HD-TV, SD-TV, Web-TV and Mobile-TV) to consumers, providers must have a video server for each IPTV service type, which increases network resource consumption. In this paper, we present a new mechanism for Multicast Broadcast Service (MBS). In particular, the proposed solution allows a provider to offer different IPTV services to varied users requests via WiMAX multihop relay access network. Results show that the proposed scheme ensures network load optimization and reduces bandwidth consumption.**

Keywords—Multihop WiMAX Relay; IEEE 802.16j; Bandwidth consumption; IPTV services; Mulitcast tree.

I. INTRODUCTION

WiMAX (Worldwide inter-operability for Microwave access) or IEEE 802.16 is communication technology used for wirelessly delivering high-speed internet service to a wide geographical zone [1]. In Metropolitan Area Networks (MANs), it is typically considered as the most reliable wireless access technology. Moreover, emerging Multihop relay (MR) wireless networks provide additional coverage or performance advantage in an access network [2]. It also provides a low-cost and flexible infrastructure that can be simultaneously utilized by multiple users for a variety of applications. However, the bandwidth and range of this wireless infrastructure make it suitable to support QoS constraints required by applications, like providing data, telecommunications (VoIP) and IPTV.

Internet Protocol Television (IPTV) is gaining recognition as a viable alternative for the delivery of video by video streaming providers [3]. The mobile IPTV technology enables users to transmit and receive multimedia traffic including television signals, video, audio, text and graphic services through IP-based wireless networks with not only full support of service quality but also with a quality experience (QoE), and security, mobility, and interactive functions [4].

Even though there are many advantages for using the IP based mobile WiMAX networks there are also some challenges. Due to the high quality IPTV services, it is impossible to guarantee the sufficient amount of the limited mobile WiMAX bandwidth for the mobile IPTV services each and every time. A Service Level Agreement (SLA) [5]-[6] between the mobile IPTV service provider and mobile WiMAX network operator in order to reserve sufficient bandwidth for the IPTV calls can increase the satisfaction level of the mobile IPTV users. For inequality between mobile WiMAX network capacity and bandwidth required by non-IPTV services and IPTV services, some requested mobile IPTV calls are blocked and some ongoing IPTV calls are dropped or quality is degraded.

WiMAX Multihope Relay technology is the adequate technology to provide IPTV services for heterogeneous user requests or their devices, because it supports QoS based multicasting functionality [7]. However, at the moment to provide varied IPTV services, providers transfer several copies of the same video content (channel), one copy for each video stream service (HD-TV, SD-TV, Web-TV and Mobile-TV). This causes increases in: bandwidth consumption and therefore influence provider video channel offering. To address this problem, we propose a solution whose objective is to reduce bandwidth consumption and supports varied services.

This paper is organized as follows. Section II presents an overview of the proposed solution. In section III, we look at the related work and background of the mechanism used. The proposed mechanism is presented in Section IV. In Section V, we present the simulation results for the proposed scheme. Finally, we conclude our paper in Section IV.

II. PROPOSED SOLUTION

In this paper, we are interested in IPTV (internet protocol television), which is an application which is currently increasing. Over the course of the next few years, the number of global IPTV subscribers is expected to grow from 28 million in 2009 to 83 million in 2013[8]. Depending on the network architecture of the service provider, it is possible to have varied services for the same video content (HD-TV, SD-TV, Web-TV and Mobile-

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

TV). Today, providers' video server architecture model is relatively simple and easy to manage. This is because providers use one server for each IPTV video service as is shown in Fig. 1. But, this increases bandwidth consumption when providers send each IPTV service flow separately.

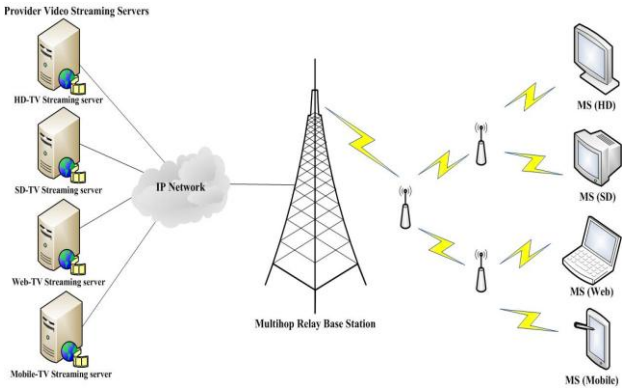


Figure 1. Traditional network architecture for IPTV services.

To solve this problem, we propose a new multicast mechanism for IPTV application over WiMAX multihop relay network, which would make it possible to reduce bandwidth consumption while satisfying QoS requirements. The main idea of our solution is to use only one video streaming server for all IPTV services as is shown in Fig. 2. We use Scalable Video Coding 'SVC' to extract video to different IPTV services. Furthermore, we propose a new multicast tree construction method by introducing MT-CID (Multicast Tunnel CID) and Transmission Identity (TxId). The TxId is attributed for each IPTV services (HD-TV, SD-TV, Web-TV and Mobile-TV) to extract video contents towards its destination with the requested quality.

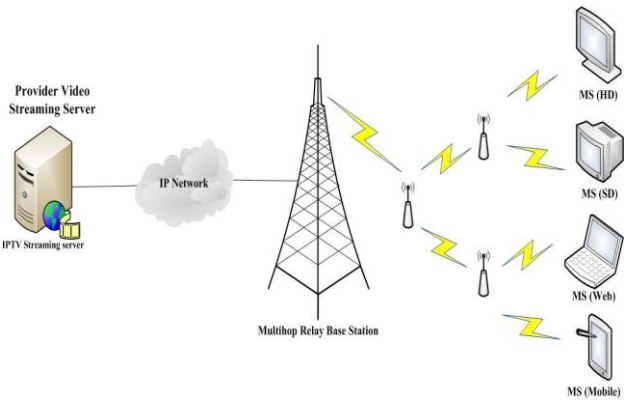


Figure 2. Proposed network architecture for IPTV services.

III. RELATED WORK

Several references have examined real time applications and more precisely IPTV in WiMax networks. For example in [7], an IGMP proxy is proposed to send Leave/Join/Report to the upstream router on behalf of the wireless sub-network. The proposal can save power consumption caused by asynchronous IGMP Query messages and improves the uplink throughput. In [4], the authors propose a SLA negotiation procedure for mobile IPTV users over mobile WiMAX networks. The Bandwidth Broker controls the allocated bandwidth for IPTV and non-IPTV users. The proposal dynamically reserves bandwidth for the IPTV services and increases the IPTV user's satisfaction level.

In [9], He et al. have proposed a standard-based cost-effective solution in order to support MBS services in WiMAX multi-hop relay network. They define a BS-oriented source-routing protocol to automatically discover relay network topology where the mobile relay station forms an ad hoc topology. They have used IGMP snooping protocol on the BS to automatically track the MBS group membership and service activation.

To address the optimal routing and bandwidth provisioning problems for survivable multicast in networks supporting IPTV services. Network-coding-based approaches and two tree-based approaches are formulated by integer linear programming in [10].

In [11], the invention relates to a multicast router of a content distribution system and a associated method adapted for receiving an upper level of a network multiplexed; in scalable video compression encoded and television video stream ,this in hand . On the other hand, it adapts the output video content to the allowed bandwidth of the lower level of the network. The authors in [11], do not take into account the constraints of QoS required by users. However, in our proposal we adapt the video stream to the user's requests while reducing bandwidth.

IV. PROPOSED IPTV MULTICAST MECHANISM

A. *Scalable Video Coding*

D Traditional digital video transmission is based on H.222.0 MPEG-2 systems for broadcasting services over satellite, cable, and wireless transmission channels, or on H.320 for conversational video conferencing services. These channels are typically characterized by a fixed spatio-temporal format of the video signal (SDTV or HDTV or CIF for H.320 video telephone) [12].

Scalable video coding (SVC) allows a single data stream to contain multiple speeds and resolutions. Solving the problem with applications that stutter, skip and crash when they cannot keep up with bit rates. SVC enables a single encoder to create a video bitstream that contains several bitstreams which can be separately decoded by dropping packets to down-sample for lower spatial resolution, lower temporal resolution or a lower quality, or a combination of the three, as required for specific client viewing hardware [13].

In the proposed mechanism, we use Scalable Video Coding to adapt the data size to the changes in video stream parameters. SVC is a highly attractive solution to the problems posed by the characteristics of modern video transmission systems [12]. In our solution, we integrate SVC functionalities for each RS in order to extract IPTV video stream to various format (e.g., HD-TV, SD-TV, etc).

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

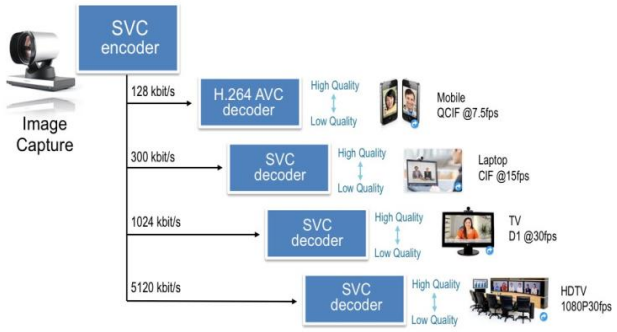


Figure 3. The Scalable Video Coding (SVC) principle [13].

B. *Multicast in WiMAX Multihop Relay*

MBS (Multicast Broadcast Service) in WiMAX is based on the ability of the WiMAX network to provide flexible and efficient mechanisms to send common content to multiple users sharing the same radio resources [9]. To support consumer's heterogeneity in WiMAX multihop relay technology, we propose a new multicast solution which is based on SVC and Transmission Identity. We attribute a Transmission Identity (TxId) for each IPTV service, as it is shown in the Table I. TxId enables BS and RS to transfer only one copy of IPTV video stream with high quality.

TABLE I. IPTV SERVICE TRANSMISSION IDENTITY MAPPING

IPTV Service	Transmission Identity
High definition (HD-TV)	TxId=1
Standard definition (SD-TV)	TxId=2
Web-TV	TxId=3
Mobile-TV	TxId=4

In this context, our intention is to propose a new multicast tree construction strategy that reduces bandwidth consumption by sending only one copy of IPTV video stream, supporting various IPTV services through the use of SVC and TxId.

C. *Multicast Tree Construction*

As mentioned on IEEE 802.16j standards, the multicast traffic will be transmitted from the MAC layer. Since, the multicast traffic will have been transmitted from the ASN-GW or other multicast [9]. A capable network which uses the IP network layer must use a mapping table to address the traffic transmission problem. Table II shows an example of mapping tables.

TABLE II. IP MULTICAST ADDRESS TO MULTICAST CID MAPPING

IP Multicast Address	MCID
224.0.0.100	MCID1
224.0.0.200	MCID2

In WiMAX MR Networks, it is necessary for MR-BS to control and manage all RSs at the same time. Compared to unicasting identical control messages are required for every RS, the use of multicasting control message by MR-BS to RSs is more efficient. In the proposed mechanism, we perform multicasting along a tunnel by using the MT-CID (Multicast Tunnel CID) which allows identifying a single link between two RS's or an RS and the MR-BS. With this solution, we can achieve multicasting for IPTV delivery along tunnel connection with less bandwidth consumption.

As 802.16j networks are comprised of multihop paths between the MR-BS and MS, routing and path management issues then arise. Although routing in such systems is tree based, there can be decisions to be made regarding which RS a particular MS should be associated with. To create paths, MR-BS makes centralized computation for the path between the MR-BS and an access RSs for both the uplink and downlink direction. After building up the path information to the destination MS, MB-RS create or update an information table that contains the mapping between a MCID and one given path. The MR-BS selects a path to carry the traffic for the new connection, and informs all the RSs on the path of the binding between the path-ID and the supported CIDs by sending a DSA-REQ message to all the RSs on the specified path [2]. Fig. 4 shows an example of Multihop relay WiMAX network by introducing MT-CID, when we have one MR-BS and five RSs.

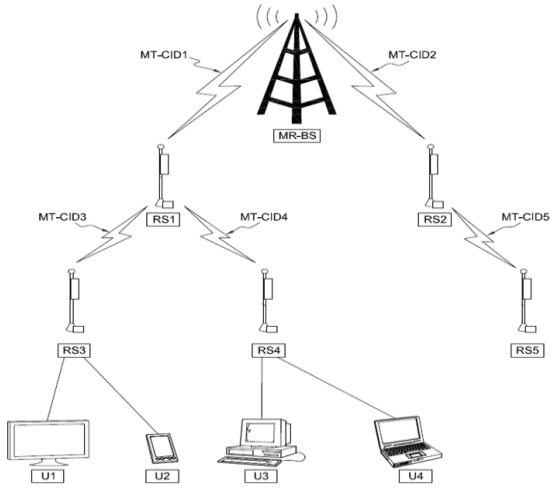


Figure 4. Multihop relay WiMAX network example

For MBS (Multicast Broadcast Service), MR-BS may initiate a multicast tree construction. When a MS wants to demand a MBS, it sends a DSA-REQ message to MR-BS to specify that it wants the MBS [14]. The procedures for establishing multicast tree are below.

When a MR-BS initiates a MBS or receives a MBS request from a MS, it verifies whether the requested MBS has been created. If not, the MR-BS creates a multicast tree for this MBS and allocates a multicast CID (MCID) to it. The MR-BS also determines the path(s) to carry this multicast service flow, attributes the adequate Transmission Identity (TxId) for each MS IPTV video stream request. In the case of our network example shown in Fig. 4, MR-BS creates paths as is illustrated in Table III, thereafter it creates the mapping between the determined path and the MT-CID like is shown in Table IV.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

TABLE III. RELAYS BELONGING TO A PATH

Path-ID	Relay
Path-ID1	RS1, RS3
Path-ID2	RS1, RS4
Path-ID3	RS2, RS5

TABLE IV. MULTICAST TUNNEL CID WITH PATH-ID MAPPING

MT-CID	Path-ID
MT-CID1	Path-ID1, Path-ID2
MT-CID2	Path-ID3
MT-CID3	Path-ID1
MT-CID4	Path-ID2
MT-CID5	Path-ID3

The MR-BS saves all information in its diffusion table. To inform all RSs with which quality (TxId), they must extract and transfer video steam content. However, MR-BS sends a DSA-REQ to indicate to RSs on the path of the binding between the path-ID, MCID and the TxId. Each RS along the path stores this information for sending multicast video stream with the appropriate quality (TxId). The MR-BS adds this path to the multicast tree that may consist of multiple paths.

If the multicast tree has been created and an MCID has been allocated to this MBS, the MR-BS would determine the path to carry this multicast service flow. Also, if the path is already in the multicast tree, the MR-BS updates its diffusion table and informs all RSs of accomplishing this change. When the parameters for a multicast service flow change, an MR-BS or MS may also send a DSC-REQ message to update these changes. All the RSs in the multicast tree of the MBS are informed of these changes.

When an MS needs to leave the multicast service, the MS sends a DSD-REQ to the MR-BS to request deleting it from the MBS. MR-BS may remove the path from a multicast tree. When an MS needs to leave the MBS, the MR-BS determines whether the path can be removed from the multicast tree. If no more MSs use this path for the MBS, the path may be removed from the multicast tree. Otherwise, the path would not be removed from the multicast tree. If the path is removed from the multicast tree, the MR-BS removes the binding between the path-ID and the MCID, it updates its diffusion table and informs all RSs in the correspond path.

In WiMAX IPTV system, when the MS selects a TV program, in addition to sending IGMP join message, it must establish a multicast connection with a MCID. However, MR-BS determines the various parameters of the requested TV flow. In order to attribute the adequate Transmission Identity to this video flow which will be used during forwarding process when a RS will send the same video stream via the same MT-CID, it would be based on the TxId. In this case, the RS sends only one copy of the video stream with the highest quality (TxId) requested by the neighbors RS's. We apply this procedure for all IPTV multicast sessions, so to support consumers' heterogeneity and to better manage the consumption of resources. Taking this into consideration, RS ensures video extraction towards all required qualities by using SVC (Scalable Video Coding).

To better understand the proposed solution, we use the topology shown in Fig. 4 for an example application. If we assume that, initially, MR-BS diffusion table is empty and the user U1 request MCID1 video stream with HD quality (TxId=1), MR-BS creates a first line:

TABLE V. DIFFUSION TABLE (A)

Path-ID	MCID	MR-BS & Relays
Path-ID1	MCID1	MR-BS(1), RS1(1), RS3 (1)

In the diffusion table above, the number in parenthesis reflects the quality of the stream (TxId) to be transmitted to the next RS in the list (or broadcast to the terminals in the case of access relay).

If afterwards the user U2 requests MCID2 video stream flow with Mobile-TV quality (TxId=4), MR-BS adds a second row to its diffusion table which is then:

TABLE VI. DIFFUSION TABLE (B)

Path-ID	MCID	MR-BS & Relays
Path-ID1	MCID1	MR-BS(3), RS1(3), RS3 (3)
Path-ID1	MCID2	MR-BS(4), RS1(4), RS3 (4)

Then, if the user U3 requests MCID1 video stream with SD quality (TxId=2), MR-BS notes from the first row of the diffusion table, the flow MCID1 is already transmitted in HD quality from MR-BS to RS1 via MT-CID1 tunnel. At the same time as shown in Table IV, the path to the user U4 borrows also MT-CID1 tunnel. Thus, it is not necessary to send MCID1 video stream flow also with SD quality via MT-CID1. The MR-BS adds a third row to diffusion table, indicating the highest quality to MCID1 flow between MR-BS and RS1, which is to say the Transmission Identity TxId = 1. The diffusion table therefore reads as follows:

TABLE VII. DIFFUSION TABLE (C)

Path-ID	MCID	MR-BS & Relays
Path-ID1	MCID1	MR-BS(1), RS1(1), RS3 (1)
Path-ID1	MCID2	MR-BS(4), RS1(4), RS3 (4)
Path-ID2	MCID1	MR-BS**(1)**, RS1(2), RS4(2)

In the diffusion table above, the TxId indicates the transmission quality of MCID1 stream to U3 user, the MT-CID1 tunnel is listed in bold to highlight this factor.

Similarly, if the user U4 requests MCID2 video stream flows with Web-TV quality (TxId=3), MR-BS finds that MCID2 flow has already transmitted with Mobile-TV quality from MR-BS to RS1. Thus, MR-BS updates the diffusion table by adding a new line to requested flow and changing the second line to indicate the highest quality:

TABLE VIII. DIFFUSION TABLE (D)

Path-ID	MCID	MR-BS & Relays
Path-ID1	MCID1	MR-BS(1), RS1(1), RS3 (1)
Path-ID1	MCID2	MR-BS**(3)**, RS1(4), RS3 (4)
Path-ID2	MCID1	MR-BS**(1)**, RS1(2), RS4(2)
Path-ID2	MCID2	MR-BS(3), RS1(3), RS4 (3)

Finally, after each, diffusion table update, MR-BS informs RSs to update their forwarding table.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

D. *Forwarding Process*

In our mechanism, the RS forwarding table contains three columns, the multicast flow identifier (MCID), the transmission identity (TxId) that represents transmission quality (ITPV service) and the tunnel MT-CID in which the identified flow must be sent. For the above example application, RS1 relay create its forwarding table as shown in Table IX.

In general, the forwarding table contains p entries, where p is the number of multicast sessions (MCID). When an intermediate RS receives a multicast packet, it first extracts a multicast session address (MCID), then the corresponding transmission identity from its forwarding table. Second, for each entry in the forwarding table, the algorithm F is executed to decide via which MT-CID tunnel video stream packets will then be transmitted.

Algorithm F

```
Begin
for (i=1; i< p; i++;)
   if ( RS forwarding table TxId = Packet header
TxId)then
        Send packets;
   else
        SVC extraction;
        Send packets;
   endif
end
```

TABLE IX. RS1 FORWARDING TABLE

MCID	TxId	MT-CID
MCID1	1	MT-CID3
MCID1	2	MT-CID4
MCID2	4	MT-CID3
MCID2	3	MT-CID4

V. SIMULATION RESULT

We performed simulations to evaluate the performance of the proposed multicast mechanism using Matlab, and compared it with the traditional mechanism of sending one copy of each video stream service, which requires four IPTV video servers. Contrary to our solution, we used a single IPTV streaming video server. We assume that the number of subscribers and IPTV channel requests are assigned randomly. We suppose that multihop WiMAX network capacity is greater than 256 Mbps. Table X recapitulates the simulation configurations.

TABLE X. SIMULATION PARAMETERS

Parameter	value	
Subscriber IPTV services request	Random	
Maximum number of IPTV channels	60	
Channels request	Random	
Maximum allowable bandwidth for the IPTV services [Mbps]	256	
IPTV services	Random [1, 4]	
Requested bandwidth by a IPTV service [Mbps]	HD-TV	6
	SD-TV	2.5
	Web-TV	0.650
	Mobile-TV	0.350

To compare the proposed mechanism with the traditional approach, we developed several scenarios based on bandwidth metrics.

Fig. 5 compares bandwidth consumption for the proposed mechanism and traditional mechanism. We assume that subscribers request IPTV channels until maximum allowable bandwidth. Fig. 5 shows that, the proposed mechanism can reduce bandwidth consumption by sending only one copy of video stream. Hence, our solution can provide the same services but it requires less bandwidth and only a single video server. When the number of subscribers or the number of IPTV channels increased the proposed mechanism gives the best performance.

Fig. 6 shows that when all IPTV channels are requested with all qualities (60 in our simulation), with the proposed mechanism there is no more bandwidth consumption. This means that we do not need the use of more bandwidth. However, in the traditional mechanism the bandwidth consumption continues to increase to provide all IPTV services.

In Fig. 7, we change the maximum number of IPTV channels to 20, in order to show that with the proposed mechanism bandwidth consumption quickly becomes constant due to all IPTV services requested and served before. On the other hand, bandwidth consumption of the traditional mechanism always increases.

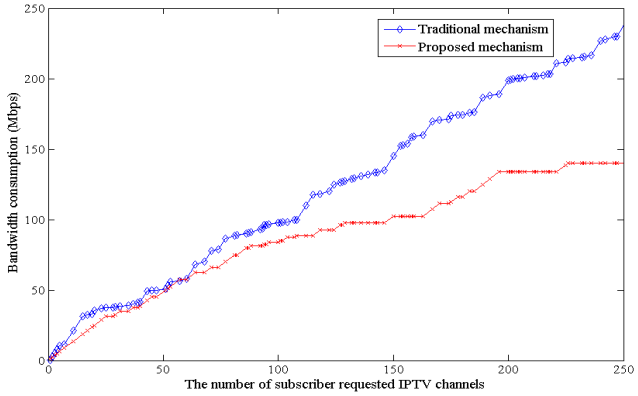


Figure 5. Average bandwidth consumption in 802.16j zone when the maximum number of IPTV channels = 60

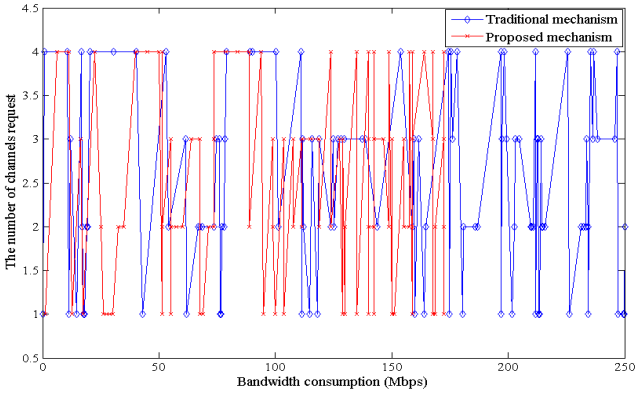


Figure 6. Average subscriber bandwidth consumption

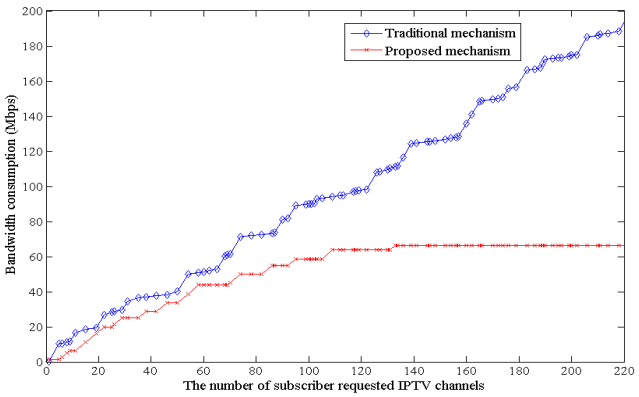


Figure 7. Average bandwidth consumption in 802.16j zone when the maximum number of IPTV channels = 20.

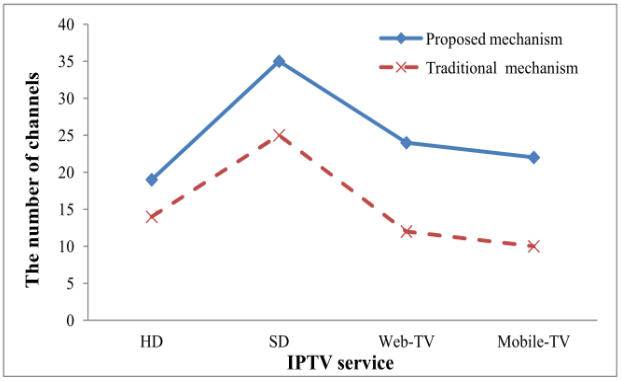


Figure 8. The average number of IPTV channels.

In Fig. 8, we further compare the number of IPTV channels that the provider could offer with the same allowable bandwidth (512 Mbps). Results show that the number of channels provided by the proposed mechanism is greater than those with the traditional mechanism for each IPTV service. This means that we have more bandwidth, so providers can offer more IPTV channels.

VI. CONCLUSION

In this work, we have addressed bandwidth consumption problems for IPTV applications. Also, we have proposed a new multicast mechanism to support several IPTV services in WiMAX multihop relay networks. Results have shown that with the proposed mechanism significant gains can be made to enhance the multicast data routing, in particular bandwidth consumption. For construction of trees, our solution has introduced MT-CID (Multicast Tunnel CID) which maps the relay paths to form a MBS tree in 802.16j MAC layer to support MBS services. By using MT-CID approach and Scalable Video Coding, RSs will only need to transfer one copy of video stream content through each tunnel which reduces network load. In addition our solution, supports various IPTV video stream services while reducing bandwidth consumption. Furthermore, it allows providers to offer more IPTV channels, rather than with only a single video server. The next challenge will be to adapt our solution to a new scheduling algorithm. We will also propose a connection admission control (CAC) mechanism to efficiently manage the resources among existing and new flows.

ACKNOWLEDGMENT

This research project is funded by France Telecom-Orange R&D and MIPS-GRTC laboratory at university of Haute Alsace.

REFERENCES

[1] Chung-Wei Lin, Yu-Cheng Chen and Ai-Chun Pang, “A New Resource Allocation Scheme for IEEE 802.16-based Networks”, 3rd IEEE VTS Asia Pacific Wireless Communications Symposium (AWPCS 2006), Aug, 2006.

[2] IEEE Std 802.16j™-2009 Part 16: “Air Interface for Fixed and Mobile Broadband Wireless Access Systems”, Multihop Relay Specification, 13 May 2009.

[3] F. E. Retnasothie, M. K. Ozdemir, T. Yiicektt, H. Celebitt, J. Zhang, and R. Muththaiah, “Wireless IPTV over WiMAX: Challenges and Applications”, In Proc. of Wireless and Microwave Technology Conference (WAMICON), Dec, 2006, Pp. 1-5.

[4] M.Z. Chowdhury, B.M. Trung, Y.M. Jang, Y. Kim, and W. Ryu, “Service Level Agreement for the QoS Guaranteed Mobile IPTV Services over Mobile WiMAX Networks”, journal CoRR, May, 2011, abs/1105.4431.

[5] D. C. Verma, “Service Level Agreements on IP Networks”, In Proc. of the IEEE, Sept, 2004, pp. 1382 - 1388.

[6] J. Sommers, P. Barford, N.G. Duffield, and A. Ron, “Multiobjective monitoring for SLA compliance”, presented at IEEE/ACM Trans. Netw., 2010, pp.652-665.

[7] Ning Liao, Yuntao Shi, Jianfeng Chen and Jun Li, “Optimized Multicast Service Management in a Mobile WiMAX TV System”, Consumer Communications and Networking Conference, Jan, 2009, pp. 1 - 5.

[8] International Television Expert Group, http://www.international-television.org/tv_market_data/ - global-iptv-forecast-2009-2013.html

[9] Chengxuan He, Oliver Yang and GuoQiang Wang, “Performance Evaluation of Multicast Routing Protocol and MBS Service Architecture in WiMAX Multi-Hop Relay Environment”, Future Networks: Cross-Layer design, April, 2008, Ottawa, Ontario, Canada.

[10] Lee S.S.W, Chen A and Po-Kai Tseng, “Optimal routing and bandwidth provisioning for survivable IPTV multicasting using network coding”, Consumer Communications and Networking Conference (CCNC), Jan, 2011 IEEE, pp. 771 – 775.

[11] Leprovost Yann and Sayadi Bessem, “Multicast router, distribution system, network and method of a content distribution”, European Patent Application, Alcatel Lucent April 2009: EP2046041.

[12] Heiko Schwarz, Detlev Marpe and Thomas Wiegand, “Overview of the Scalable Video Coding Extension of the H.264/AVC Standard”, IEEE Transactions On Circuits And Systems For Video Technology, VOL. 17, NO. 9, Sept, 2007, pp. 1103 - 1120.

[13] http://nextgenlog.blogspot.com/2010/09/freescale-dsp-tackles-scalable-video.html

[14] Hang Liu and Mingquan Wu, “Connection Management for Multicast and Broadcast Services (MBS)”, IEEE C802.16j-07/272, 2007-04-05.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Oriented 2-hop Forwarding Approach on Voids Boundaries in Wireless Sensor Networks

Mohamed Aissani, Sofiane Bouznad, Abdelmalek Hariza, and Salah-Eddine Allia

Research Unit in Computer Science (UERI), Ecole Militaire Polytechnique (EMP)
P.O. Box 17, Bordj-El-Bahri 16111, Algiers, Algeria
{maissani, bouznad.sofiane}@gmail.com
{malik-abd, s.alia}@hotmail.com

Abstract — **Scalable geographical routing protocols suffer from voids that appear in Wireless Sensor Networks (WSNs). Several techniques are proposed in literature to handle this problem, but they present some limits, particularly in time-critical applications. Consequently, we propose in this paper a new 2-hop forwarding approach that orients any packet which arrives at a boundary node in the shortest path towards the sink. The handled voids can be either closed within a deployed WSN or open on the network boundary. To keep unchanged the actual size of a void for a long time, the use of a 2-hop forwarding mode is privileged to preserve the limited energy of boundary nodes. The information needed for our approach is provided by simple and reactive algorithms that we propose in this paper to discover and maintain the boundaries of voids. Associated with the SPEED real-time routing protocol, our proposal performs very well in terms of packet delivery ratio, control packet overhead, network and boundary nodes energy consumption.**

Keywords–Sensor networks; geographical routing; closed voids; open voids; void-handling techniques.

I. INTRODUCTION

WSNs can be deployed quickly in sensitive and/or difficult to access areas. Their mission is usually to monitor an area, to take regular measurements and to send alarms to the sink(s) of the deployed network. Many applications using WSNs are then emerging in several areas, such as defense, security, health, agriculture and smart homes. They generally used geographical routing ensuring scalability and allowing positive progression of packets towards the sink. However, geographical routing has two major problems. First, it is not applicable if a sender node has no opportunity to know its geographical locations. This problem can be solved by virtual coordinate systems. Second, there may be voids between a source node and a sink. These voids can be concave, convex, closed or open. Conversely to the closed voids that appear within a deployed WSN, the open voids are frequently formed on the boundary of this network. A geographical routing path towards the sink can be failed due to lack of relay nodes because of a void.

To handle the problem of voids in geographical routing, several solutions are proposed in literature [1-14], but they present some shortcomings, particularly in case of time-critical applications using WSNs. As a contribution in resolving this problem, we propose an oriented 2-hop forwarding approach handling effectively all kinds of voids in WSNs. To do so, we also propose four reactive algorithms to discover and then maintain each void that appear in a deployed WSN. Then each data packet received by a boundary node is forwarded towards its destination by using the shortest path and the minimum number of boundary nodes. This strategy aims to reduce the packet end-to-end delay, to economize the energy of boundary nodes and then to preserve for a long time the actual form of each discovered void. Note that the present work improves our previous work [13, 14] by handling both open and closed voids, using a 2-hop forwarding mode on the void boundary and maintaining dynamically each discovered void in a WSN.

The rest of the paper is organized a follows. Section II presents the problem of voids and discuses the existing void-handling techniques. Section III provides two efficient algorithms for discovery and maintenance of voids in WSNs. Section IV proposes an oriented 2-hop forwarding mode to use by each boundary node. Section V evaluates performance of the proposed approach. Section VI concludes the paper.

II. VOID PROBLEM IN GEOGRAPHICAL ROUTING

Routing voids are areas where nodes cannot forward data packets or completely unavailable. These voids are formed due to either the random deployment of nodes or the node failure because of various reasons, such as circuit failure, destruction or energy exhaustion. Therefore, packets to forward are often blocked in their positive progression towards their destination.

Suppose the example in Figure 1, where black nodes are boundary nodes and node s has to forward data packets to destination d. Node s is stuck because it has no neighbor so close to d to be selected as a forwarder node; i.e., the FS (Forwarding candidate neighbors Set) of node s is empty. Once received by node s, data packets cannot progress positively towards destination d. Thanks to a recovery mode, those packets will be forwarded to node j (or to node k) in a negative progression to bypass the void. This scenario, called the local minimum phenomenon, often occurs when a void appears in a WSN. We then say that s is a stuck (or a blocked) node.

Without using an adequate void-handling technique, data packets can be removed in a WSN wasting the nodes resources and communications can be lost between some pairs of nodes. Such behavior is undesirable in a time-critical application because the loss of some captured information can interfere with the network mission. To reduce the negative impact of voids on the effectiveness of geographical routing, void-handling techniques are available in literature. They fall into two classes: those based on the right-hand rule [1-6] and those using the backpressure rule [9-12].

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

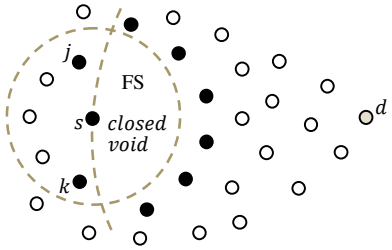


Figure 1. The void problem: the FS of sender s towards destination d is empty.

The techniques belonging to the first class use boundary nodes to route a stuck packet. In most cases, they use long recovery paths, especially in the case of open voids. Proposed in [1], the GPSR algorithm uses two forwarding modes: the greedy mode and the perimeter mode. When a sender node is not blocked, it forwards the current packet to the closest neighbor to the destination node (i.e.; greedy mode). As a result, the destination is approached hop by hop until reached by the packet. When the greedy mode fails, the packet is routed by using a face routing (i.e.; a perimeter forwarding on a planar graph) to bypass the void met. The right-hand rule is thus used on the void boundary until the packet reaches the closest node to the destination. Several other algorithms using the face routing were proposed later [2-5]. However, it has been shown in [15] that the use of planarization algorithms, such as Gabriel graphs [1], reduces the number of useful links in a WSN. This influences the exploration of multiple routing paths allowing load balancing, link-failure tolerance and network fluidity. This is not tolerable in WSNs dedicated to time-critical applications.

However, the techniques belonging to the second class uses the backpressure messages, that are broadcasted by stuck nodes near a void, to route the next packets in alternative paths. He et al. [9] describes SPEED; a QoS routing protocol providing a soft end-to-end real-time to all flows routed in a WSN. In this protocol, each node updates information on its neighbors and uses geographical routing to select paths. In addition, SPEED aims to ensure a certain delivery speed so that each application can estimate the packet end-to-end delay. It deals with a void as it handles a permanent congestion. When a packet is stuck, the sender node drops the packet and broadcasts a backpressure message informing its neighbors about the void met. Then the stuck node will not be considered by the neighbors in their future routing decisions. When neighbors of a node are all stuck, the actual packet is dropped and a backpressure message is broadcasted. This process is repeated until an alternative route is found or the source node is reached by the successive backpressure messages. Extensions to the SPEED protocol have been proposed later in [10-12], but the void-avoidance scheme of the protocol was not modified in these extensions.

Indeed, the right-hand rule is not effective in bypassing voids, especially in case of open voids. It requested a lot of boundary nodes and often used long paths on voids boundaries, resulting in excessive energy consumption of boundary nodes and delays packets due to the overload of these bypassing paths. Then the voids tend to expand rapidly due to energy depletion, complicating the sensor network mission. Similarly, the backpressure rule generates many control packets and removes data packets at stuck nodes in concave areas of some voids. Consequently, routing paths become long because of multiple backtrackings which overload links and delay packets. These packets might be removed in the sensor network after expiration of their deadline. This is again not desirable for time-critical applications. To overcome these weaknesses, we propose in this paper an efficient 2-hop forwarding approach that orients correctly towards the sink each packet received by a boundary node. The proposed approach uses two new mechanisms: the first one, is called OVA-vb (Oriented Void Avoidance on a closed void boundary), which handles the closed voids within the network whereas the second one, is called OVA-nb (Oriented Void Avoidance on the network boundary), and it handles open voids on the network boundary.

Note that the closed voids in a WSN are discovered by the VBD (Void-Boundary Discovery) algorithm and maintained by the VBM (Void-Boundary Maintenance) algorithm that we propose in the next section.

III. PROPOSED VBD AND VBM ALGORITHMS

Existing algorithms for discovery and maintenance of voids, such as BOUNDHOLE [6] and other algorithms based on the right-hand rule [7, 8], insert information on boundary nodes of a void in the VD (Void Discovery) packet, increasing both memory and energy requirements of these nodes and then reducing scalability. In addition, these algorithms perform a periodical check of a void and rediscovers the entire void if one boundary node fails, or it would be economic to discover locally only the changed segment. BOUNDHOLE [6] does not distinguish between an open void and a closed one. Indeed, the outside of a deployed WSN is considered as a great void and data packets that stuck on the network boundary will go on long bypassing paths. Also, the algorithms using the right-hand rule to discover a void do not consider an open void as a particular problem to be handled and they only discover the voids located inside the network. To alleviate these shortcomings, we propose below two effective algorithms (VBD and VBM). The VBD algorithm identifies all nodes forming the boundary of a closed void, calculates and then communicates the void information (i.e.; center and radius) to each discovered boundary node. The VBM algorithm detects and then updates any changes that occur on the boundary of a closed void that was already discovered in a WSN.

A. *Proposed VBD algorithm*

To discover the boundary nodes of a closed void, the VBD algorithm uses the right-hand rule on a Gabriel graph (GG) which preserves the network connectivity [1]. This graph is formed by neighbors of a boundary node where intersections between edges are eliminated to avoid loops. The VBD algorithm operates in initial, intermediate and final phases.

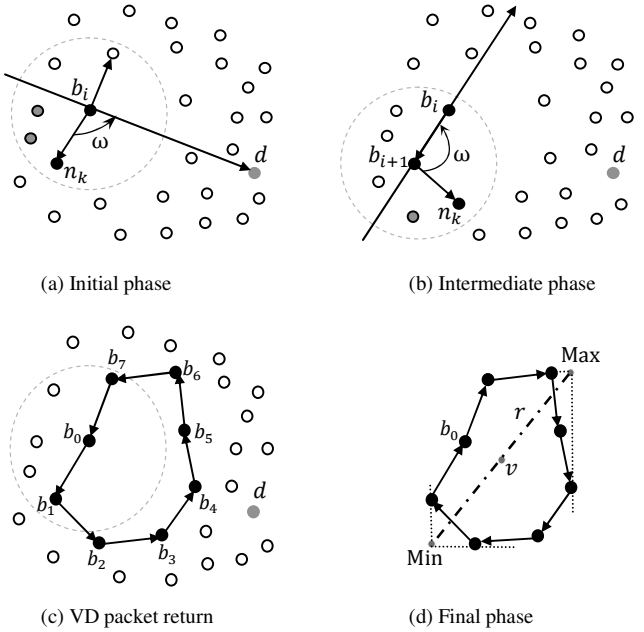
1) Initial phase: when a blocking situation is detected (i.e., FS=ϕ), node b_i performs the following tasks: (a) broadcasts a 1-hop VP (Void back-Pressure) packet announcing its non-availability for the time VT (Void Time-discovery), (b) drops the data packet to increase the network fluidity and (c) sends a

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

VD (Void-boundary Discovery) packet, marked by its ID, to next boundary-neighbor n_k located at right of vector $\overrightarrow{b_i d}$ (i.e., node n_k having the smallest ω shown in Figure 2-a).

2) Intermediate phase: when receiving the VD packet, the boundary node b_{i+1} broadcasts a VP packet and sends the VD packet to the next intermediate boundary neighbor n_k located at right of $\overrightarrow{b_{i+1}b_i}$ as shown in Figure 2-b. This process is repeated by each intermediate neighbor ($b_{i+2}, b_{i+3}, \ldots$) until the VD packet will be received by the initiator boundary node b_0 at the end of its trip around the void (Figure 2-c).

3) Final phase: by receiving the VD packet at the end of its trip, node b_0 performs the following tasks: (a) extracts from the VD packet the points Min and Max of the discovered boundary $\{b_0, b_1, \ldots, b_n\}$, (b) calculates center v of the void which is the midpoint of the segment $\overline{\text{Min Max}}$, and its radius r which is given by: $r = Distance(\text{Min}, \text{Max})/2$, (c) drops the VD packet and then (d) sends a VU (Void-boundary Update) packet, marked by its ID, through the discovered boundary in the opposite direction of the VD packet (Figure 2-d).



(a) Initial phase (b) Intermediate phase

(c) VD packet return (d) Final phase

Figure 2. The void discovery process in the VBD algorithm.

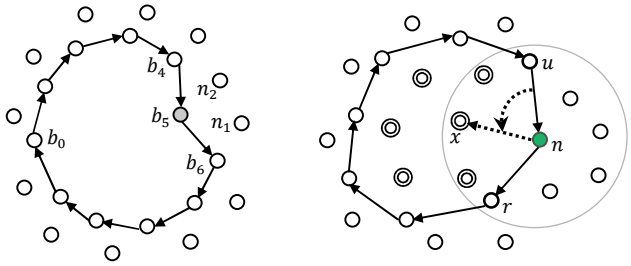
Note that before forwarding the VD packet, node b_i updates its field V1Up by the ID of its successor n_k and checks the field NodeUp in the VD packet. If this field identifies a neighbor then b_i updates its field V2Down (2-hop downstream boundary node) by NodeUp, else V2Down is updated by V1Down. Similarly, each node b_i that receives a VU packet updates its fields about the void and checks the field NodeUp in the VU packet. If this field identifies a neighbor then b_i updates its field V2Up by NodeUp, else V2Up receives V1Up. Note that the fields V2Up (2-hop upstream boundary node) and V2Down are used by the 2-hop forwarding mode of the OVA-vb mechanism which reduces both the node energy consumption and the packet end-to-end delay.

B. Proposed VBM algorithm

Some boundary nodes of a closed void in a WSN may stop working for various reasons. Also, new nodes can be deployed within a closed void to repair it. Thus the proposed VBM algorithm handles these two situations as follows.

1) Boundary-node failure: each boundary node b_i can detect the absence of its direct ascendant boundary neighbor b_{i-1} thanks to its field V1Up. When b_{i-1} expires in the neighbors table T of node b_i, the later discovers a new segment of nodes and connects it to the old segment of the void by running the VBD algorithm. When node b_5 fails in Figure 3-a, node b_6 discovers the new segment of nodes $b_6 n_1 n_2 b_4$ that connects to the old segment $b_4 b_0 b_6$ of the void. When the two segments are connected, the VD packet continues its trip to bring the full information about the new boundary of the closed void. Upon receiving the VD packet at the end, node b_i (i.e.; node b_6 in Figure 3-a) runs the final phase of the VBD algorithm updating the void information in fields of the boundary nodes.

2) Deployment of nodes within a closed void: by receiving a location beacon from a new neighbor x, boundary node n checks if x is located inside the void. Based on its updated fields V1Up and V1Down, node n uses its 1-hop boundary neighbors u and r (Figure 3-b) to execute the following verification: if $\widehat{unx} < \widehat{unr}$ then x is located inside the void. If so, node n sends a VS (Void Suppression) packet, marked by its ID, to visit the boundary of the repaired void. Upon receiving the VS packet, each boundary node removes from its list of voids (VList) the repaired void. Note that parts of a void may still exist due to repairing process, but they will be met later by packets and then discovered by the VBD algorithm.



(a) Node b_5 fails. (b) New node x sends a beacon.

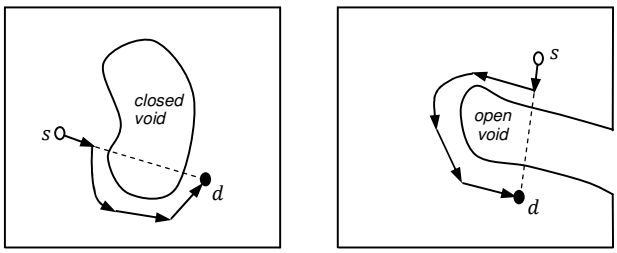
Figure 3. The void-maintenance cases in the VBM algorithm.

IV. PROPOSED 2-HOP FORWARDING APPROACH

The proposed 2-hop forwarding approach aims to orient towards the sink any packet that arrives at a boundary node by using an optimal path, as shown in Figure 4. When a sender node s has to forward a packet p towards destination d, it forms its FS (Forwarding candidate neighbors Set) and then distinguishes the three following cases: 1) sender s has no information about voids, 2) sender s is on the network boundary and 3) sender s is on the boundary of a closed void.

1) Sender s has no information about voids ($s.VList=\phi$): if FS is empty then sender s runs the VBD algorithm to discover the void met, else it forwards packet p to its neighbor n in FS

(i.e., one of the hatched nodes in Figure 5). The forwarder n is selected according to the protocol routing metric, such as the relay speed used in SPEED [9].



(a) OVA-vb mechanism (b) OVA-nb mechanism

Figure 4. Packet orientation at a boundary node in our approach.

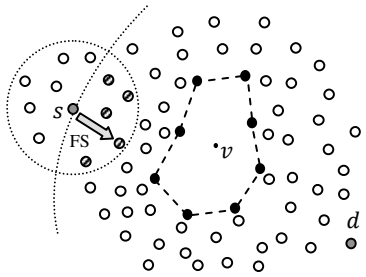
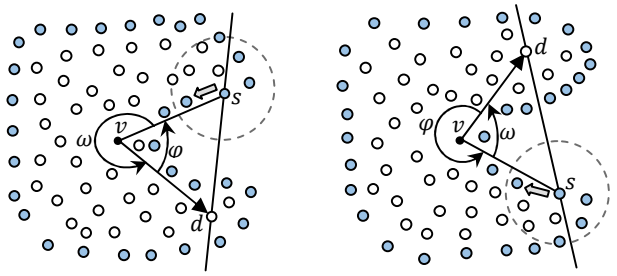


Figure 5. Case 1: sender s has not information about voids.

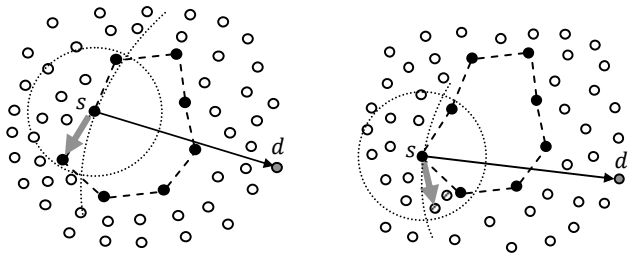
2) Sender s is on the network boundary ($s.NBorder=1$): the sender s uses the OVA-nb mechanism that we proposed in [17] to orient p towards its destination node d by using a 2-hop forwarding mode on the network boundary. Thus, sender s uses the angles $\varphi = \widehat{dvs}$ and $\omega = \widehat{svd}$ (Figure 6) to select the next forwarder n. If φ<ω (Figure 6-a) then sender s selects n from its neighbors located at the right of line (sd), else (Figure 6-b) n is selected from the neighbors of s that are located at the left of line (sd). More details about OVA-nb are given in [17].



(a) (b)

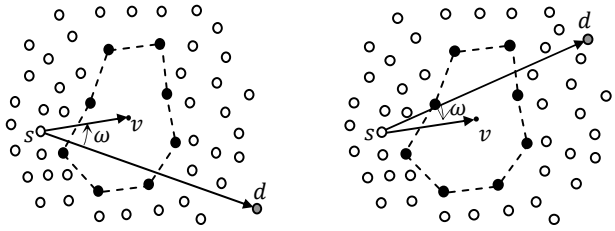
Figure 6. Case 2: sender s is on the network boundary [17]. The next forwarder is located right (a) or left (b) of line (sd).

3) Sender s is on boundary of a closed void ($s.VBorder=1$): the sender s uses the OVA-vb mechanism based on a 2-hop forwarding mode on the void boundary. Thus, packet p is oriented in the correct direction around the void by using a non-boundary node as next forwarder as soon as possible, to preserve the actual form of the void for a long time. If sender s have to route on the void boundary (Figure 7-a), it forwards p to its 2-hop upstream node identified by V2Up (or 2-hop downstream node identified by V2Down) depending on the packet orientation (i.e., right or left of $\overrightarrow{sv}$). If not (i.e., there is at least one non-boundary node in FS as shown in Figure 7-b), sender s forwards p to a neighbor n selected from its RFS (reduced FS) which is formed by the hatched nodes in Figure 7-b. The selection of n is made according to the implemented protocol metric, such as the relay speed used in SPEED [9]. Note that to orient p arround a closed void, sender s uses the angle ω shown in Figure 8. If $sin(\omega) > 0$ (Figure 8-a) then the packet orientation must be at right of $\overrightarrow{sv}$ (i.e., $p.\text{Orient}=1$). If not (Figure 8-b) then the packet orientation must be at left of $\overrightarrow{sv}$ (i.e., $p.\text{Orient}=0$). By using the field Orient in p, sender s forms its RFS by neighbors in FS that are located either at right of $\overrightarrow{sd}$ when $p.\text{Orient}=1$ or at left of $\overrightarrow{sd}$ when $p.\text{Orient}=0$.



(a) (b)

Figure 7. Case 3: sender s is on the boundary of a closed void.



(a) Orientation at right of $\overrightarrow{sv}$ (b) Orientation at left of $\overrightarrow{sv}$

Figure 8. Packet orientation updating in the OVA-vb mechanism.

Note that any changes that occur on the boundary (or inside) of a closed void will be immediately detected by a boundary node and then updated by this later after running the VBM algorithm. The reactive maintenance of the open voids on the network boundary is guaranteed by the NBM algorithm that we proposed in [17].

V. Performance Evaluation

To evaluate performance of the proposed 2-hop forwarding approach, we associate the proposed OVA-vb and OVA-nb mechanisms with the well-known SPEED real-time routing protocol by using the ns-2 simulator [16]. We compare the resulting protocol, called SPEED-vb, with the GPSR and SPEED traditional protocols. Note that to handle voids SPEED uses the backpressure rule and GPSR the right-hand rule. The parameters used in our simulations are given in TABLE I.

We used a terrain (scene) with a size of 800m×800m and 960 deployed nodes. For each simulation, we create a void in

the center of this terrain with a radius varying between 60m and 200m. Six sources selected randomly from the left side of the void generate periodic CBR packets to the first destination placed at right side of this void. Meanwhile, six other sources selected randomly from the right side of the void generate periodic CBR packets to the second destination placed at the left side of the same void. The rate of the sources is set to 1 packet/second and the desired delivery speed (the $S_{setpoint}$ defined in [9]) is set to 600m/s, which leads to an end-to-end packet deadline of 100ms. Each point in our graphs is the average of 15 simulations carried out in the same conditions, but with different sources selected randomly for each simulation. To measure the routing performance with the presence of congestion, two nodes located below the void exchanged packets with a rate of 10 packets/second during the simulation time which is set to 224 seconds.

TABLE I. SIMULATION PARAMETERS.

MAC Layer	IEEE 802.11
Radio Layer	RADIO-NONOISE
Propagation Model	TwoRayGround
Antenna Model	OmniAntenna
Queue Model	Queue/DropTail/PriQueue
Queue Size	50 paquets
Transmission chanal	WirelessChannel
Wireless Interface	WirelessPhy
Bandwidth	200 Kb/s
CBR Packet Size	32 octets
Energy Model	energyModel de ns-2
Communication Range	40 m
Transmission Power	0.666 w
Reception Power	0.395 w

We evaluate performance of protocols SPEED-vb, SPEED and GPSR. We vary the void radius and we measure the packet delivery ratio, the control packet overhead, the network and the boundaries energy consumption per delivered packet. The figures 9, 10, 11 and 12 show that the protocols' performance decreases each time the void radius grows because they use long paths around the void. Therefore, deadline of many packets expires before reaching their destination and then they are dropped in the network because we suppose a time-critical application. We also note that the proposed SPEED-vb protocol is the most efficient with the presence of both small and large voids in a WSN. This is due to the performance of the proposed mechanisms used by the boundary nodes.

Figure 9 shows that SPEED is the worst protocol in delivering packets, especially when a void radius is greater than 120m. This protocol overloads its upstream nodes by the backpressure messages generation near the voids. Following the spread of these messages, some sources are blocked and many packets are removed when their deadline expires in congested links. For an acceptable packet deadline (100ms), GPSR performs better than SPEED tanks to its face routing scheme used by boundary nodes. GPSR generates less control packets (Figure 11) that reduces the network congestion. With the adequate orientation of packets ensured by the proposed mechanisms, the SPEED-vb protocol uses the shortest and smoother routing paths compared to the SPEED and GPSR protocols. Therefore, the packet delivery ratio achieved by SPEED-vb is the highest (Figure 9).

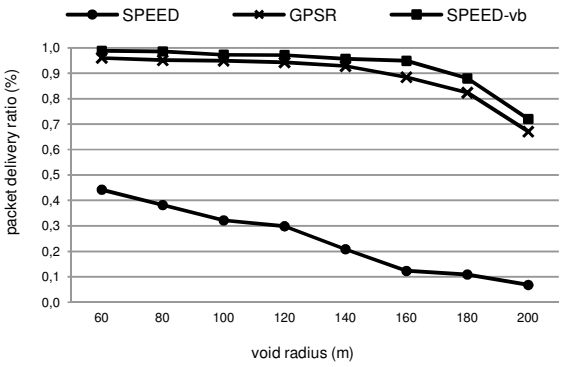


Figure 9. Packet delivery ratio *vs.* Void radius.

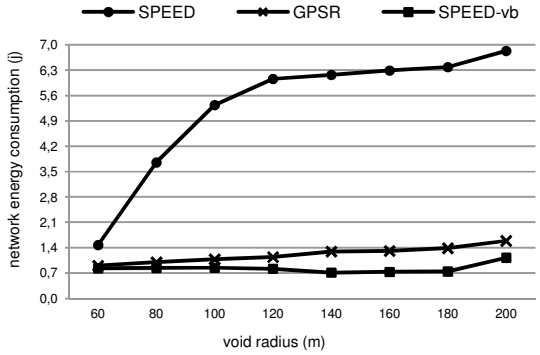


Figure 10. Network energy consumption per delivered packet.

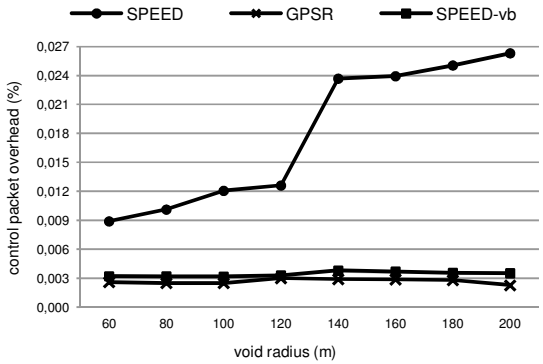


Figure 11. Control packet overhead.

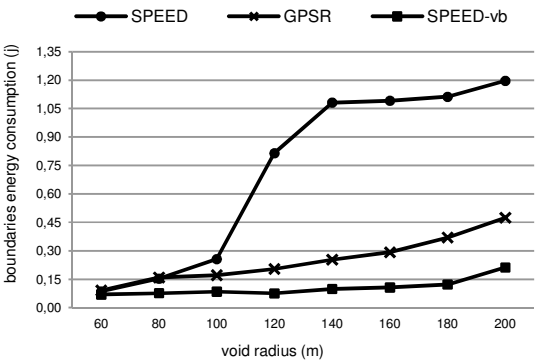


Figure 12. Boundaries energy consumption per delivered packet.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

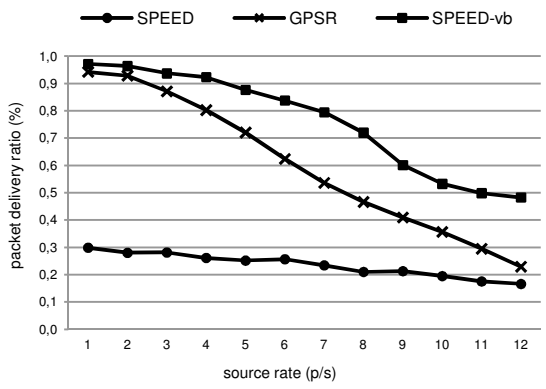


Figure 13. Packet delivery ratio *vs.* Source rate.

For some delivered packets, SPEED consumes much energy of both the network (Figure 10) and the boundary nodes (Figure 12). This is due to excessive control packets generated by SPEED and its useless routing of delayed packets in the network. GPSR is more efficient than SPEED in term of network energy consumption, but it consumes more energy of boundary nodes, especially when the void radius exceeds 100m (Figure 12). For these large voids, GPSR routes most packets on the long parts of the boundary. In the other hand, our SPEED-vb protocol achieves the best tradeoff between the packet delivery ratio and the energy consumption (Figure 10). Since GPSR always uses a unique path connecting a source to the sink, it does not achieve a good node energy balancing. Figure 13 shows these limits when the rate exceeds 3 p/s and a void with 120m as radius is created in center of the terrain. In the other hand, SPEED-vb delivers many data packets thanks to its void-handling mechanisms.

VI. CONCLUSION AND FUTURE WORK

We have proposed an oriented 2-hop forwarding approach that provides to each packet received by a boundary node the shortest path towards the sink. Our void-tolerant approach uses two complementary mechanisms: the first one handles the open voids located on the network boundary and the second one handles the closed voids located within the network. These mechanisms use simple and reactive algorithms that we have proposed to discover and then to maintain each void that appears in a deployed WSN. We have associated them with the well-known SPEED routing protocol, designed for real-time applications, and the resulting protocol, called SPEED-vb, achieved the best performance compared to the traditional GPSR and SPEED protocols. The SPEED-vb protocol was able to respond to the shortcomings of the existing void-handling techniques, which are based either on the right-hand rule, such as GPSR, or on the backpressure rule, such as SPEED.

Since we are interested by time-critical applications based on WSNs, our future work will focus on the sequencing of data packets at a node based on the time remaining to reach the sink. The objective is to reduce the number of removed critical packets due to deadline expiration. We also plan to check how our idea can be applied to congested regions and or to the voids created due other problems, like intermittent connectivity.

REFERENCES

[1] B. Karp and H. Kung, "GPSR: Greedy perimeter stateless routing for wireless networks," Proc. of the IEEE Conference on Mobile Computing and Networking, pp. 243-254, Boston, USA, August 6-11, 2000.

[2] L. Moraru, P. Leone, S. Nikoletseas, and J.D.P Rolim, "Near optimal geographical routing with obstacle avoidance in wireless sensor networks by fast-converging trust-based algorithms," Proc. of the 3rd ACM Workshop on Q2SWinet, pp. 31–38, Greece, October 22-26, 2007.

[3] L. Moraru, P. Leone, S. Nikoletseas, and J. Rolim, "Geographical routing with Early Obstacles Detection and Avoidance in Dense Wireless Sensor Networks," Lecture Notes in Computer Science, Vol. 5198(1), pp. 148-161, 2008.

[4] F. Kuhn, R. Wattenhofer, and A. Zollinger, "An Algorithmic Approach to Geographical routing in Ad Hoc and Sensor Networks," IEEE/ACM Transactions on Networking, Vol. 16(1), pp. 51–62, 2008.

[5] F. Huc, A. Jarry, P. Leone, L. Moraru, S. Nikoletseas, and J. Rolim, "Early Obstacle Detection and Avoidance for All to All Traffic Pattern in Wireless Sensor Networks," Lecture Notes in Computer Science, Vol. 5804(1), pp. 102–115, 2009.

[6] Q. Fang, J. Gao, and L.J. Guibas, "Locating and bypassing routing holes in sensor networks," Journal of Mobile Networks and Applications, Vol. 11(2), pp. 187–200, October 2006.

[7] F. Yu, Y. Choi, S. Park, E. Lee, Y. Tian, M. Jin, and S.H. Kim, "Anchor node based virtual modeling of holes in wireless sensor networks," Proc. of the International Conference on Communications (ICC), pp. 3120-3124, Beijing, China, May 19-23, 2008.

[8] [10]F. Yu, S. Park, E. Lee, and S.H. Kim, "Hole Modeling and Detour Scheme for Geographical routing in Wireless Sensor Networks," Journal of Communications and Networks, Vol. 11(4), pp. 327-336, Aug. 2009.

[9] T. He, J.A. Stankovic, C. Lu, and T. Abdelzaher, "A Spatiotemporal Communication Protocol for Wireless Sensor Networks," IEEE Transactions on Parallel and Distributed Systems, Vol. 16(10), pp. 995-1006, October 2005.

[10] E. Felemban, C.G. Lee, and E. Ekici, "MMSPEED: Multipath Multi-SPEED Protocol for QoS Guarantee of Reliability and Timeliness in Wireless Sensor Networks," IEEE Transactions on Mobile Computing, Vol. 5(6), pp. 738-754, June 2006.

[11] W. Cheng, L. Yuan, Z. Yang, and X. Du, "A real-time Routing Protocol with Constrained Equivalent Delay in Sensor Networks," Proc. of the 11th IEEE Symposium on Computers and Communications (ISCC), pp. 597-602 Italy, June 26-29, 2006.

[12] L. Zhao, B. Kan, Y. Xu, and X. Li, "FT-SPEED: A Fault-Tolerant Real-Time Routing Protocol for Wireless Sensor Networks," Proc. of the International Conference on WiCom, Shanghai, pp. 2531–2534, China, September 21-25, 2007.

[13] M. Aissani, A. Mellouk, N. Badache, and B. Saidani, "Oriented Void Avoidance Scheme for Real-Time Routing Protocols in Wireless Sensor Networks," Proc. of the IEEE GLOBECOM Conference, pp. 83-87, New Orleans, LA, USA, 30 Nov. - 04 Dec. 2008.

[14] M. Aissani, A. Mellouk, N. Badache, and M. Boumaza, "A Novel Approach for Void Avoidance in Wireless Sensor Networks," International Journal of Communication Systems (IJCS), Vol. 23(8), pp. 945–962, 2010.

[15] K. Seada, A. Helmy, and R. Govindan, "Modeling and analyzing the correctness of geographic face routing under realistic conditions," Ad-Hoc Networks, Vol. 5(6), pp. 855–871, August 2007.

[16] Collaboration between researchers at UC Berkeley, LBL, USC/ISI, and Xerox PARC, "The ns Manual", *http://www.isi.edu/nsnam/ns/*, last consultation in June 2011.

[17] M. Aissani, S. Bouznad, A. Hariza, and S.E. Allia, "An effective mechanism for handling open voids in wireless sensor networks," Proc. of the 5th International Conference on Sensor Technologies and Applications (SENSORCOMM), pp. 24-29, Riviera, France, August 21-27, 2011.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Fequency Offset Estimation for OFDM Systems in Non-Gaussian Noise Channels

Changha Yu, Jong In Park, Youngpo Lee, and Seokho Yoon†
College of Information and Communication Engineering
Sungkyunkwan University
Suwon, South Korea
e-mail: {dbckdgk, pji17, leeyp204, and †syoon}@skku.edu
†Corresponding author

***Abstract*—In this paper, the frequency offset estimation schemes robust to the non-Gaussian noise for orthogonal frequency division multiplexing (OFDM) systems are addressed. First, a maximum-likelihood (ML) estimation scheme in non-Gaussian noise is proposed, and then a simpler estimation scheme based on the ML estimation scheme is presented. Numerical results show that the proposed schemes offer robustness and a substantial performance improvement over the conventional estimation scheme in non-Gaussian noise channels.**

Keywords-frequency offset estimation; maximum-likelihood; non-Gaussian noise; OFDM; training symbol

I. INTRODUCTION

Due to its immunity to multipath fading and high spectral efficiency, orthogonal frequency division multiplexing (OFDM) has been adopted as a modulation format in a wide variety of wireless systems such as digital video broadcasting-terrestrial (DVB-T), wireless local area network (WLAN), and worldwide interoperability for microwave access (WiMAX) [1]-[4]. However, the OFDM is very sensitive to the frequency offset (FO) caused by Doppler shift or oscillator instabilities, and thus, the frequency offset estimation is one of the most important technical issues in OFDM systems [1], [5]. Specifically, we are concerned about the FO estimation based on training symbols, which provides a better performance than that based on the blind approach [5].

Conventionally, the FO estimation schemes have been proposed under the assumption that the ambient noise is a Gaussian process [6]-[8], which is generally justified with the central limit theorem. However, it has been observed that the ambient noise often exhibits non-Gaussian nature in wireless channels, mostly due to the impulsive nature originated from various sources such as car ignitions, moving obstacles, lightning in the atmosphere, and reflections from sea waves [9], [10]. The conventional estimation schemes developed under the Gaussian assumption on the ambient noise could suffer from severe performance degradation under such non-Gaussian noise channels.

In this paper, we propose robust FO estimation schemes in non-Gaussian noise channels. First, we derive a maximum-likelihood (ML) FO estimation scheme in non-Gaussian noise modeled as a complex isotropic Cauchy noise, and then, derive a simpler estimation scheme with a lower complexity. From numerical results, the proposed schemes are confirmed to offer a substantial performance improvement over the conventional scheme in non-Gaussian noise channels.

The rest of this paper is organized as follows. Section II introduces the related works on the FO estimation, and the signal model is described in Section III. In Section IV, two FO estimation schemes are proposed for OFDM systems in non-Gaussian noise environments. Section V demonstrates the numerical results. Section VI concludes this paper.

II. RELATED WORKS

Several schemes [6]-[8] have been proposed to estimate the FO of OFDM signals assuming the Gaussian noise environments. The FO estimation scheme in [6] uses a training symbol with two identical halves to estimate the FO within the sub-carrier spacing. Then, using the other training symbol containing a pseudonoise (PN) sequence, the scheme corrects the remaining FO that is a multiple of the sub-carrier spacing. The scheme in [7] uses the best linear unbiased estimation (BLUE) principle requiring only one training symbol with more than two identical parts. Moreover, its estimation performance is quite close to the Cramer-Rao lower bound (CRLB). In [8], joint ML FO estimation scheme was derived when the training symbol is repeated multiple times. Specifically, the scheme in [8] exploits the correlation of any pair of repetition patterns providing optimized performance in the OFDM systems.

III. SIGNAL MODEL

The kth OFDM sample $x(k)$ is generated by the inverse fast Fourier transform (IFFT), and can be expressed as

$$x(k) = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} X_m e^{j2\pi km/N}, \qquad (1)$$

for $k = 0,1,\cdots,N-1$, where X_m is a phase shift keying (PSK) or quadrature amplitude modulation (QAM) symbol in the mth subcarrier and N is the size of the IFFT. Then, the cyclic prefix (CP) of the OFDM symbol is inserted, whose length is generally designed to be longer than the channel impulse response, to avoid the intersymbol

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

interference (ISI). Assuming that the timing synchronization is perfect, we can express the kth received OFDM sample $r(k)$ after removing the CP as

$$r(k) = \sum_{l=0}^{L-1} h(l)x(k-l)e^{j2\pi k\varepsilon/N} + n(k) \quad (2)$$

for $k = 0,1,\cdots,N-1$, where $h(l)$ is the lth channel coefficient of a multipath channel with length L, ε is the FO normalized to the subcarrier spacing $1/N$, and $n(k)$ is the kth sample of additive noise.

In this paper, we adopt the complex isotropic symmetric α stable (CISαS) model for the independent and identically distributed noise samples $\{n(k)\}_{k=0}^{N-1}$ this model has been widely employed due to its strong agreement with experimental data [11], [12]. The probability density function (pdf) of $n(k)$ is then given by [11]

$$f_n(\rho) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\gamma\left(u^2+v^2\right)^{\frac{\alpha}{2}} - j\Re\{\rho(u-jv)\}} dudv, \quad (3)$$

where $\Re\{\cdot\}$ denotes the real part, the dispersion $\gamma > 0$ is related to the spread of the pdf, and the characteristic exponent $\alpha \in (0,2]$ is related to the heaviness of the tails of the pdf: A smaller value of α indicates a higher degree of impulsiveness, whereas a value closer to 2 indicates a more Gaussian behavior.

A closed-form expression of (3) is not known to exist except for the special cases of $\alpha = 1$ (complex isotropic Cauchy) and $\alpha = 2$ (complex isotropic Gaussian). In particular, we have

$$f_n(\rho) = \begin{cases} \dfrac{\gamma}{2\pi}\left(|\rho|^2+\gamma^2\right)^{-\frac{3}{2}}, & \text{when } \alpha = 1 \\ \dfrac{1}{4\pi\gamma}\exp\left(-\dfrac{|\rho|^2}{4\gamma}\right), & \text{when } \alpha = 2. \end{cases} \quad (4)$$

Due to such a lack of closed-form expressions, we concentrate on the case of $\alpha = 1$: We shall see in Section V that the estimation schemes obtained for $\alpha = 1$ are not only more robust to the variation of α, but they also provide a better performance for most values of α, than the conventional estimation scheme.

IV. Proposed Schemes

A. Maximum-likelihood FO Estimation Scheme

In estimating the FO, we consider a training symbol $\{x(k)\}_{k=0}^{N-1}$ with two identical halves as in [6], i.e., $x(k) = x(k+N/2)$ for $k = 0,1,\cdots,N/2-1$. From (2), we have

$$r(k+N/2) - r(k)e^{j\pi\varepsilon} = n(k+N/2) - n(k)e^{j\pi\varepsilon} \quad (5)$$

for $k = 0,1,\cdots,N/2-1$. Observing that $n(k+N/2) - n(k)e^{j\pi\varepsilon}$ obeys the complex isotropic Cauchy distribution with dispersion 2γ (since the distribution of $-n(k)e^{j\pi\varepsilon}$ is the same as that of $n(k)$, and assumed that the noise samples of CISαS model are independent as in [13]), we obtain the pdf

$$f_{\mathbf{r}}(\mathbf{r}\,|\,\varepsilon) = \prod_{k=0}^{\frac{N}{2}-1} \frac{\gamma}{\pi\left(\left|r(k+N/2) - r(k)e^{j\pi\varepsilon}\right|^2 + 4\gamma^2\right)^{\frac{3}{2}}} \quad (6)$$

of $\mathbf{r} = \{r(k+N/2) - r(k)e^{j\pi\varepsilon}\}_{k=0}^{N/2-1}$ conditioned on ε. The ML estimation is then to choose $\hat{\varepsilon}$ such that

$$\begin{aligned} \hat{\varepsilon} &= \arg\max_{\tilde{\varepsilon}} [\log f_{\mathbf{r}}(\mathbf{r}\,|\,\tilde{\varepsilon})] \\ &= \arg\min_{\tilde{\varepsilon}} \Lambda(\tilde{\varepsilon}), \end{aligned} \quad (7)$$

where $\tilde{\varepsilon}$ denotes the candidate value of ε and the log-likelihood function $\Lambda(\tilde{\varepsilon}) = \sum_{k=0}^{N/2-1} \log\left\{\left|r(k+N/2) - r(k)e^{j\pi\tilde{\varepsilon}}\right|^2 + 4\gamma^2\right\}$ is a periodic function of $\tilde{\varepsilon}$ with period 2: The minima of $\Lambda(\tilde{\varepsilon})$ occur at a distance of 2 from each other, causing an ambiguity in estimation. Assuming that ε is distributed equally over positive and negative sides around zero, the valid estimation range of the ML estimation scheme can be set to $-1 < \varepsilon \le 1$, as in [6]. The estimation scheme (7) will be called the Cauchy ML estimation (CMLE) scheme.

B. Low-complexity FO Estimation Scheme

The CMLE scheme is based on the exhaustive search over the whole estimation range ($|\varepsilon| \le 1$), which requires high computational complexity. Thus, we propose a low-complexity FO estimation scheme with the reduced set of the candidate values.

In order to obtain the reduced set of the candidate values, we exploit the property that $\varepsilon = 1/\pi\angle\{x^*(k)x(k+N/2)\} = 1/\pi\angle\{r^*(k)r(k+N/2)\}$ in the absence of noise. Based on this property, we obtain the set of the candidate values

$$\bar{\varepsilon}(k) = \frac{1}{\pi}\angle\{r^*(k)r(k+\frac{N}{2})\}, \text{ for } k = 0,1,\cdots,\frac{N}{2}-1. \quad (8)$$

Exploiting the set of the candidate values in (8), the FO estimate $\hat{\varepsilon}_L$ can be obtained as follows

$$\hat{\varepsilon}_L = \arg\min_{\bar{\varepsilon}(k)} \Lambda(\bar{\varepsilon}(k)), \quad \text{for} \quad k = 0,1,\cdots,\frac{N}{2}-1. \quad (9)$$

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

In the following, (9) is denoted as the low-complexity CMLE (L-CMLE) scheme. Using only $N/2$ candidate values, the L-CMLE scheme can offer an almost same performance as the CMLE scheme with the exhaustive search, the performance verified by numerical results in Section V.

V. NUMERICAL RESULTS

In this section, the proposed CMLE and L-CMLE schemes are compared with the Gaussian ML estimation (GMLE) scheme in [6] in terms of the mean squared error (MSE) by computer simulations using Matlab program and computational complexity. We assume the following simulation parameters: The IFFT size $N = 64$, FO $\varepsilon = 0.25$, length 8 samples of CP, the interval of search spacing 0.001 for the CMLE scheme, and a multipath Rayleigh fading channel with length $L = 8$ and an exponential power delay profile of $E[|h(l)|^2] = \exp(-l/L)/\{\sum_{l=0}^{L-1}\exp(-l/L)\}$ for $l = 0,1,\cdots,7$, where $E[\cdot]$ denotes the statistical expectation. Since CIS α S noise with $\alpha < 2$ has an infinite variance, the standard signal-to-noise ratio (SNR) becomes meaningless for such a noise. Thus, we employ the geometric SNR (GSNR) defined as $\frac{E[|x(k)|^2]}{4C^{-1+2/\alpha}\gamma^{2/\alpha}}$, where $C = \exp\{\lim_{m\to\infty}(\sum_{i=1}^{m}\frac{1}{i} - \ln m)\} \simeq 1.78$ is the exponential of the Euler constant [14]. The GSNR indicates the relative strength between the information-bearing signal and the CIS α S noise with $\alpha < 2$. Clearly, the GSNR becomes the standard SNR when $\alpha = 2$. Since γ can be easily and exactly estimated using only the sample mean and variance of the received samples [15], it may be regarded as a known value: Thus, γ is set to 1 without loss of generality.

Figs. 1-4 show the MSE performances of the CMLE, L-CMLE, and GMLE schemes as a function of the GSNR when $\alpha = 0.5,1,1.5,$ and 2, respectively. From the figures, we can clearly observe that the CMLE and L-CMLE schemes have a better estimation performance compared with that of the GMLE scheme for most values of α, except for $\alpha = 2$. Another important observation is that the estimation performance of the L-CMLE scheme is almost same as that of the CMLE scheme. From this observation, it is confirmed that the trial values for the L-CMLE scheme is reasonable. Numerical results show that proposed schemes not only outperform the conventional scheme in non-Gaussian noise environments, but also provide similar performance in Gaussian noise ($\alpha = 2$) environments. This can clearly explain a robustness of proposed schemes to the variation of the channel environments. In short, when the type of the noise is not known, the L-CMLE scheme can be an effective solution with robust performance to the noise.

Table I shows the computational complexity of CMLE, L-CMLE, and GMLE schemes, where S denotes the number of search spacing for the CMLE scheme. The GMLE scheme requires $(3N-2)$ real additions and $(2N+1)$ real multiplications only. On the other hand, the CMLE scheme requires $SN(3N-1)$ real additions and $SN(5N/2)$ real multiplications by choosing the most likely candidate among the SN candidates. Using $N/2$ reliable candidates only, the L-CMLE scheme reduced the number of operations to $N/2(3N-1)+N$ real additions and $(N/2+1)(5N/2)$ real multiplications.

TABLE I. COMPUTATIONAL COMPLEXITY OF THE FO ESTIMATION SCHEMES

	CMLE	**L-CMLE**	**GMLE**
Number of candidates	SN	$N/2$	-
Real additions	$3N-1$ per candidate	$3N-1$ per candidate + N	$3N-2$
Real multiplications	$5N/2$ per candidate	$5N/2$ per candidate $+5N/2$	$2N+1$

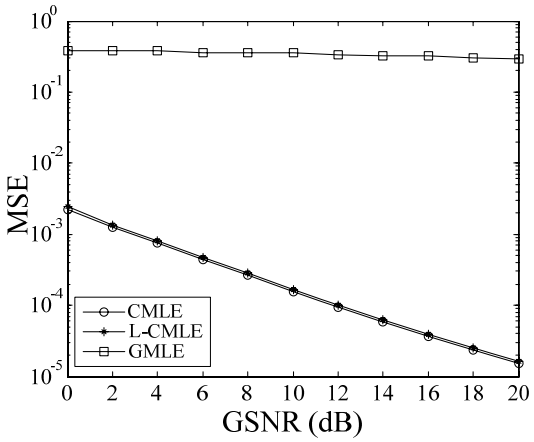


Figure 1. The MSE performances of the CMLE, L-CMLE, and GMLE schemes as a function of the GSNR when $\alpha = 0.5$.

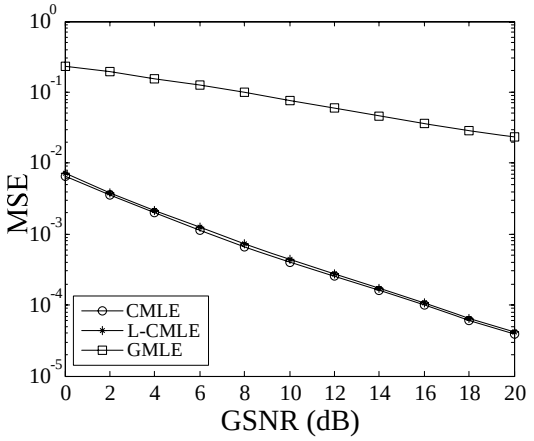


Figure 2. The MSE performances of the CMLE, L-CMLE, and GMLE schemes as a function of the GSNR when $\alpha = 1$.

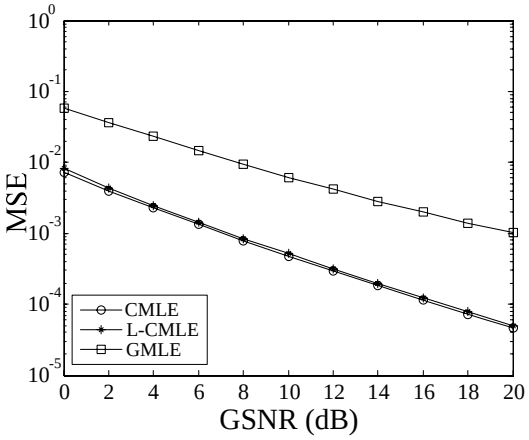


Figure 3. The MSE performances of the CMLE, L-CMLE, and GMLE schemes as a function of the GSNR when $\alpha = 1.5$.

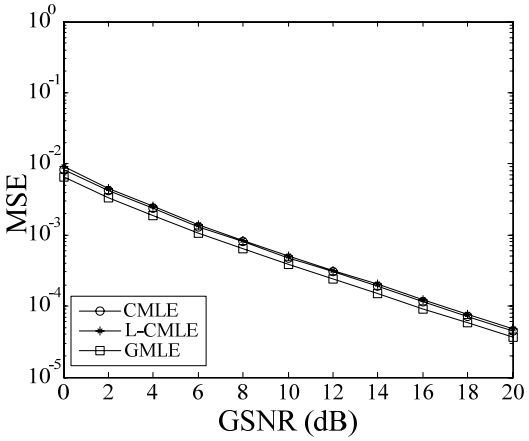


Figure 4. The MSE performances of the CMLE, L-CMLE, and GMLE schemes as a function of the GSNR when $\alpha = 2$.

VI. CONCLUSION

In this paper, we have proposed FO estimation schemes in non-Gaussian noise channels. First, an ML estimation scheme in non-Gaussian noise channel has been proposed, and then a simpler estimation scheme based on the ML estimation scheme has been presented. From the numerical results, it has been confirmed that the proposed schemes offer robustness and a substantial performance improvement over the conventional estimation scheme in non-Gaussian noise channels.

ACKNOWLEDGMENT

This research was supported by the National Research Foundation (NRF) of Korea under Grant 2011-0018046 with funding from the Ministry of Education, Science and Technology (MEST), Korea, by the Information Technology Research Center (ITRC) program of the National IT Industry Promotion Agency under Grant NIPA-2012-H0301-12-1005 with funding from the Ministry of Knowledge Economy (MKE), Korea, and by National GNSS Research Center program of Defense Acquisition Program Administration and Agency for Defense Development.

REFERENCES

[1] R. V. Nee and R. Prasad, *OFDM for Wireless Multimedia Communications*. Boston, MA: Artech House, 2000.

[2] IEEE Std. 802.11h, *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification: Spectrum and Transmit Power Management Extensions in the 5GHz Band in Europe*, IEEE, 2003.

[3] M. Morelli, C.-C. J. Kuo, and M.-O. Pun, "Synchronization techniques for orthogonal frequency division multiple access (OFDMA): a tutorial review," *Proc. IEEE*, vol. 95, no. 7, pp. 1394-1427, July 2007.

[4] A. Awoseyila, C. Kasparis, and B.G. Evans, "Robust time-domain timing and frequency synchronization for OFDM systems," *IEEE Trans. Consumer Electron.*, vol. 55, no. 2, pp. 391-399, May 2009.

[5] T. Hwang, C. Yang, G. Wu, S. Li, and G. Y. Li, "OFDM and its wireless applications: a survey," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1673-1694, May 2009.

[6] T. M. Schmidl and D. C. Cox, "Robust frequency and timing synchronization for OFDM," *IEEE Trans. Commun.*, vol. 45, no. 12, pp. 1613-1621, Dec. 1997.

[7] M. Morelli and U. Mengali, "An improved frequency offset estimator for OFDM applications," *IEEE Commun. Lett.*, vol. 3, no. 3, pp. 75-77, Mar. 1999.

[8] J.-W. Choi, J. Lee, Q. Zhao, and H.-L. Lou, "Joint ML estimation of frame timing and carrier frequency offset for OFDM systems employing time-domain repeated preamble," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 311-317, Jan. 2010.

[9] T. K. Blankenship and T. S. Rappaport, "Characteristics of impulsive noise in the 450-MHz band in hospitals and clinics," *IEEE Trans. Antennas, Propagat.*, vol. 46, no. 2, pp. 194-203, Feb. 1998.

[10] P. Torío and M. G. Sánchez, "A study of the correlation between horizontal and vertical polarizations of impulsive noise in UHF," *IEEE Trans. Veh. Technol.*, vol. 56, no. 5, pp. 2844-2849, Sep. 2007.

[11] C. L. Nikias and M. Shao, *Signal Processing With Alpha-Stable Distributions and Applications*. New York, NY: Wiley, 1995.

[12] H. G. Kang, I. Song, S. Yoon, and Y. H. Kim, "A class of spectrum-sensing schemes for cognitive radio under impulsive noise circumstances: structure and performance in nonfading and fading environments," *IEEE Trans. Veh. Technol.*, vol. 59, no. 9, pp. 4322-4339, Nov. 2010.

[13] J. Ilow and D. Hatzinakos, "Impulsive noise modeling with stable distributions in fading environments," *Proc. IEEE Signal Process. Workshop on Statistical Signal and Array Process.*, pp. 140-143, Corfu, Greece, June 1996.

[14] T. C. Chuah, B. S. Sharif, and O. R. Hinton, "Nonlinear decorrelator for multiuser detection in non-Gaussian impulsive environments," *Electron. Lett.*, vol. 36, no. 10, pp. 920-922, May 2000.

[15] X. Ma and C. L. Nikias, "Parameter estimation and blind channel identification in impulsive signal environments," *IEEE Trans. Signal Process.*, vol. 43, no. 12, pp. 2884-2897, Dec. 1995.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Unambiguous BOC Signal Acquisition Based on Recombination of Sub-Correlations

Changha Yu, Jong In Park, Youngpo Lee, and Seokho Yoon†
College of Information and Communication Engineering
Sungkyunkwan University
Suwon, Korea
e-mail: {dbckdgk, pji17, leeyp204, and †syoon}@skku.edu
†Corresponding author

***Abstract*—In this paper, we propose a novel unambiguous acquisition scheme for binary offset carrier (BOC) signals. Specifically, we first find out that the side-peaks arise due to the fact that the BOC autocorrelation is made up of the sum of the irregularly shaped sub-correlations, and then, propose an unambiguous acquisition scheme by recombining the sub-correlations. The proposed scheme is shown to remove the side-peaks completely for any type of BOC signal and to provide a better performance than the conventional scheme in terms of the incorrect acquisition probability and mean acquisition time.**

Keywords-acquisition; ambiguous; binary offset carrier; correlation function; side-peak

I. INTRODUCTION

Recently, new global navigation satellite systems (GNSSs) such as Galileo and global positioning system (GPS) modernization are being developed to satisfy the increasing demand for GNSS-based services such as location-based service (LBS) and emergency rescue service (ERS) and complement the existing GNSSs such as GPS [1]-[2]. Currently, new GNSSs are designed to use the same frequency band of the existing GNSSs: for example, the E1 and E5 bands of Galileo are overlapped with the L1 and L5 bands of GPS, respectively [1], [3]. Thus, if a Galileo signal is modulated by a conventional scheme such as phase shift keying (PSK) used in GPS, it would suffer from co-channel interference. To overcome these problems, binary offset carrier (BOC) modulation has been proposed, where a high degree of spectral separation between the BOC-modulated signals and the others is achieved by shifting the signal energy from the band center [4]. The BOC signal is generated through the product of a spreading pseudo random noise (PRN) code and a sine-phased or cosine-phased square wave sub-carrier, and denoted by $\mathrm{BOC}_{\sin}(kn,n)$ or $\mathrm{BOC}_{\cos}(kn,n)$ depending on which of the sine-phased or cosine-phased sub-carriers are used, where k and n are the ratios of the PRN code chip period to the sub-carrier period and the PRN code chip rate to 1.023 MHz, respectively [3], [5]. For larger values of k, more separated spectrums are obtained, reducing the co-channel interference more effectively.

However, the BOC signal has multiple side-peaks on both sides of the main-peak of its autocorrelation function. Moreover, the number of side-peaks increases as the value of k becomes larger. Thus, the correlation-based synchronization schemes [6]-[11] originally proposed for PRN code synchronization would suffer from the ambiguous problem in the BOC signal synchronization due to the side-peaks in the BOC autocorrelation.

Several unambiguous acquisition schemes [12]-[16] have been proposed in order to tackle the problem. In [12]-[14], sideband filtering was used to deal with the ambiguous problem in the BOC signal acquisition; however, these schemes destroy the sharpness of the main-peak of the BOC autocorrelation function, degrading the BOC signal tracking performance severely. In [15], an interesting unambiguous acquisition scheme that maintains the sharp main-peak of the BOC autocorrelation function was proposed combining the correlation between the BOC and PRN signals with the BOC autocorrelation; however, this scheme is applicable to only $\mathrm{BOC}_{\sin}(n,n)$ signals. In [16], a generalized unambiguous acquisition scheme including the scheme in [15] as a special case was proposed. This scheme is applicable to generic $\mathrm{BOC}_{\sin}(kn,n)$ signals; however, its extension to generic $\mathrm{BOC}_{\cos}(kn,n)$ signals is not straightforward. In addition, in [17], an unambiguous scheme applicable to both sine-phased and cosine-phased BOC signals was proposed by the authors preliminarily; however, the performance of the scheme becomes worse as the value of k increases.

In this paper, a novel unambiguous acquisition scheme applicable to both $\mathrm{BOC}_{\sin}(kn,n)$ and $\mathrm{BOC}_{\cos}(kn,n)$ signals is proposed based on a recombination of the sub-correlations making up the BOC autocorrelation. The scheme is found to remove the side-peaks of the BOC autocorrelation completely, while keeping the sharp shape of the main-peak. Moreover, it is demonstrated that the scheme offers a performance improvement over the scheme in [16] in terms of the incorrect acquisition probability and mean acquisition time (MAT).

The remainder of this paper is organized as follows. Section II analyzes the sub-correlations making up the BOC autocorrelation. In Section III, an unambiguous acquisition scheme with no side-peak is proposed by recombining the

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

sub-correlations. Section IV presents numerical results, and finally, Section V concludes this paper.

II. BOC SUB-CORRELATIONS

The BOC signal $b(t)$ can be expressed as

$$b(t) = \sqrt{P}\sum_{i=-\infty}^{\infty} c_i\, p_{T_c}(t - iT_c)\, d_{\lfloor iT_c/T \rfloor}(t)\, s(t), \tag{1}$$

where P is the signal power, $c_i \in \{-1,1\}$ is the ith chip of a PRN code with a period of T, T_c is the PRN code chip period, $p_{T_c}(t)$ is the PRN code waveform defined as a unit rectangular pulse over $[0, T_c)$, $d_{\lfloor iT_c/T \rfloor}(t)$ is the navigation data, where $d_x(t)$ is the xth navigation data and $\lfloor x \rfloor$ is the largest integer not larger than x, and

$$s(t) = \begin{cases} \displaystyle\sum_{u=0}^{2k-1} (-1)^u\, p_{T_s}(t - iT_c - uT_s), & \text{for } \mathrm{BOC}_{\sin}(kn, n), \\ \displaystyle\sum_{u=0}^{4k-1} (-1)^{\lceil \frac{u}{2} \rceil}\, p_{\frac{T_s}{2}}\left(t - iT_c - \frac{uT_s}{2}\right), & \text{for } \mathrm{BOC}_{\cos}(kn, n) \end{cases} \tag{2}$$

is the square wave sub-carrier, where T_s is the sub-carrier pulse duration of $T_c/2k = 1/(2kn \times 1.023\,\mathrm{MHz})$, $p_{T_s}(t)$ is the unit rectangular sub-carrier pulse waveform over $[0, T_s)$, and $\lceil x \rceil$ is the smallest integer not less than x. In this paper, focusing on the problem of ambiguity due to side-peaks, we assume that there is a pilot channel for acquisition [18] so that no data modulation is present during acquisition (i.e., $d_{\lfloor iT_c/T \rfloor}(t) = 1$ for all i), and do not consider the effect of the secondary code. Then, considering that the PRN code period T is generally much larger than the PRN code chip period T_c and the out-of-phase autocorrelation of a PRN code is designed to be as low as possible for easy signal acquisition, we can obtain the correlation (normalized to the signal power) between the received and locally generated BOC signals as [19]

$$\begin{aligned} R_{\sin}^{k}(\tau) &= \frac{1}{P}\int_0^T (b(t-\tau) + w(t))\, b(t)\, dt \\ &\simeq \sum_{u=0}^{2k-1}\left(N\sum_{v=0}^{2k-1} (-1)^{u+v}\, \Lambda_{T_s}(\tau - (u-v)T_s) + w_{\sin}^{u} \right) \end{aligned} \tag{3}$$

for $\mathrm{BOC}_{\sin}(kn, n)$ and

$$R_{\cos}^{k}(\tau) \simeq \sum_{u=0}^{4k-1}\left(N\sum_{v=0}^{4k-1} (-1)^{\lceil \frac{u}{2} \rceil + \lceil \frac{v}{2} \rceil}\, \Lambda_{\frac{T_s}{2}}(\tau - (u-v)\frac{T_s}{2}) + w_{\cos}^{u} \right) \tag{4}$$

for $\mathrm{BOC}_{\cos}(kn, n)$, where τ is the phase difference between the received and locally generated BOC signals, N is a correlation length and would be generally equal to or less than the PRN code period (normalized to T_c), $w(t)$ is the additive white Gaussian noise (AWGN) process with mean zero and one-sided noise power spectral density N_0,

$$w_{\sin}^{u} = \frac{1}{\sqrt{P}}\int_0^T \sum_{i=-\infty}^{\infty} (-1)^u\, c_i\, p_{T_s}(t - iT_c - uT_s)\, w(t)\, dt, \tag{5}$$

$$w_{\cos}^{u} = \frac{1}{\sqrt{P}}\int_0^T \sum_{i=-\infty}^{\infty} (-1)^{\lceil \frac{u}{2} \rceil}\, c_i\, p_{\frac{T_s}{2}}\left(t - iT_c - \frac{uT_s}{2}\right) w(t)\, dt, \tag{6}$$

and

$$\Lambda_x(\tau) = \begin{cases} x - |\tau|, & |\tau| \le x, \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

is the triangular function of height x and area x^2. After denoting the triangular functions and noise terms in (3) and (4) as

$$R_{\sin}^{k,u}(\tau) = N\sum_{v=0}^{2k-1} (-1)^{u+v}\, \Lambda_{T_s}(\tau - (u-v)T_s) + w_{\sin}^{u} \tag{8}$$

and

$$R_{\cos}^{k,u}(\tau) = N\sum_{v=0}^{4k-1} (-1)^{\lceil \frac{u}{2} \rceil + \lceil \frac{v}{2} \rceil}\, \Lambda_{\frac{T_s}{2}}\left(\tau - (u-v)\frac{T_s}{2}\right) + w_{\cos}^{u}, \tag{9}$$

respectively, we can re-write $R_{\sin}^{k,u}(\tau)$ and $R_{\cos}^{k,u}(\tau)$ as

$$\begin{aligned} R_{\sin}^{k,u}(\tau) &= N\sum_{v=0}^{2k-1} (-1)^{u+v}\, \Lambda_{T_s}(\tau - (u-v)T_s) + w_{\sin}^{u} \\ &= \sum_{l=0}^{N-1}\sum_{v=0}^{2k-1} (-1)^{u+v} \int_{(2kl+u-1)T_s}^{(2kl+u)T_s} p_{T_s}(t - lT_c - uT_s) \\ &\quad \cdot p_{T_s}(t - \tau - lT_c - vT_s)\, dt + w_{\sin}^{u} \\ &= \sum_{l=0}^{N-1} \frac{1}{P}\int_{(2kl+u-1)T_s}^{(2kl+u)T_s} r(t)\, b(t)\, dt, \end{aligned} \tag{10}$$

and similarly,

$$R_{\cos}^{k,u}(\tau) = \sum_{l=0}^{N-1} \frac{1}{P}\int_{(\frac{4kl+u}{2} - \frac{1}{2})T_s}^{(\frac{4kl+u}{2})T_s} r(t)\, b(t)\, dt, \tag{11}$$

where $r(t) = b(t-\tau) + w(t)$. From (10) and (11), we can see that $R_{\sin}^{k,u}(\tau)$ and $R_{\cos}^{k,u}(\tau)$ are sub-correlations making up the correlations (3) and (4), respectively, and which are shown for $k = 1$ in the absence of noise in Fig. 1. From the figure, we can see that the main-peaks of the sub-correlations are coherently combined through the summation of the sub-

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

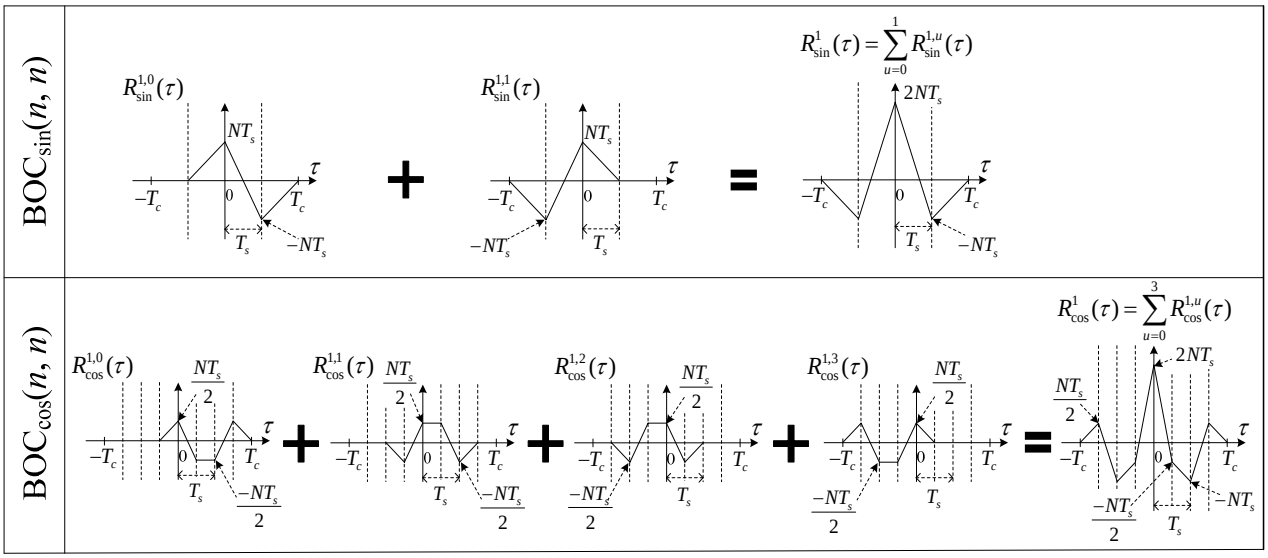


Figure 1. BOC autocorrelation and the assiciated sub-correlations for $\mathrm{BOC}_{\sin}(n,n)$ and $\mathrm{BOC}_{\cos}(n,n)$.

correlations, thus forming the sharp main-peak of the BOC autocorrelation, and on the other hand, the sub-peaks of the sub-correlations are irregularly spread around the main-peaks, and thus, the summation of the sub-correlations results in the multiple side-peaks of the BOC autocorrelation. In the next section, we propose a novel unambiguous acquisition scheme, removing the side-peaks completely through a recombination of the sub-correlations.

III. PROPOSED SCHEME

Fig. 2 shows the unambiguous correlation functions of the proposed scheme for $\mathrm{BOC}_{\sin}(n,n)$ and $\mathrm{BOC}_{\cos}(n,n)$ as an example. From the figure, we can observe that $R_{\sin}^{k,0}(\tau)$ and $R_{\sin}^{k,2k-1}(\tau)$ and $R_{\cos}^{k,0}(\tau)$ and $R_{\cos}^{k,4k-1}(\tau)$ are symmetric with respect to $\tau = 0$ and have only a single overlapped peak at $\tau = 0$ for $\mathrm{BOC}_{\sin}(kn,n)$ and $\mathrm{BOC}_{\cos}(kn,n)$, respectively. Thus, if the two sub-correlations are summed, a main-peak with a larger magnitude (than that of the main-peak of a sub-correlation) is obtained without increasing the magnitudes of the side-peaks, and on the other hand, the difference between the two sub-correlations yields side-peaks only, whose magnitudes and positions are the same as those of the side-peaks in the sum of the two sub-correlations. Thus, the difference between the two sub-correlations might be used to remove the side-peaks in the sum of the two sub-correlations, leaving only the main-peak. This observation is the key motivation of the proposed scheme.

Since the side-peaks in the sum and difference of the two sub-correlations are out-of-phase and in-phase at $\tau < 0$ and $\tau > 0$, respectively, however, we cannot remove the side-peaks in the sum of the two sub-correlations completely through the subtraction between the sum and difference of the two sub-correlations. To align the phases of the side-peaks in the sum and difference of the two sub-correlations, thus, we use the sum of the absolute values of the two sub-correlations, obtaining the side-peaks with the same slopes as those of the side-peaks in the absolute difference of the two sub-correlations. Fig. 2 shows that the subtraction of the absolute difference of the two sub-correlations from the sum of the absolute values of the two sub-correlations yields an unambiguous correlation function with a single main-peak and no side-peak.

Since the unambiguous correlation function is generated by using only two sub-correlations out of $2k\,(4k)$ sub-correlations, the height of the main-peak is limited to $2NT_s\,(NT_s)$ for $\mathrm{BOC}_{\sin}(kn,n)\,(\mathrm{BOC}_{\cos}(kn,n))$ while the height of the BOC autocorrelation is $2kNT_s$. Considering that each sub-correlation has only a small portion of the total energy, we multiply the unambiguous correlation function with the BOC autocorrelation to obtain an unambiguous correlation function with higher main-peak, allowing it possible to make use of more signal energy.

From the above discussions, the proposed unambiguous correlation function can be expressed as

$$R_{\sin}^{k,\mathrm{proposed}}(\tau) = R_{\sin}^{k}(\tau)\Big(\big|R_{\sin}^{k,0}(\tau)\big| + \big|R_{\sin}^{k,2k-1}(\tau)\big| - \big|R_{\sin}^{k,0}(\tau) - R_{\sin}^{k,2k-1}(\tau)\big|\Big) \tag{12}$$

for $\mathrm{BOC}_{\sin}(kn,n)$ and

$$R_{\cos}^{k,\mathrm{proposed}}(\tau) = R_{\cos}^{k}(\tau)\Big(\big|R_{\cos}^{k,0}(\tau)\big| + \big|R_{\cos}^{k,4k-1}(\tau)\big| - \big|R_{\cos}^{k,0}(\tau) - R_{\cos}^{k,4k-1}(\tau)\big|\Big) \tag{13}$$

for $\mathrm{BOC}_{\cos}(kn,n)$.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

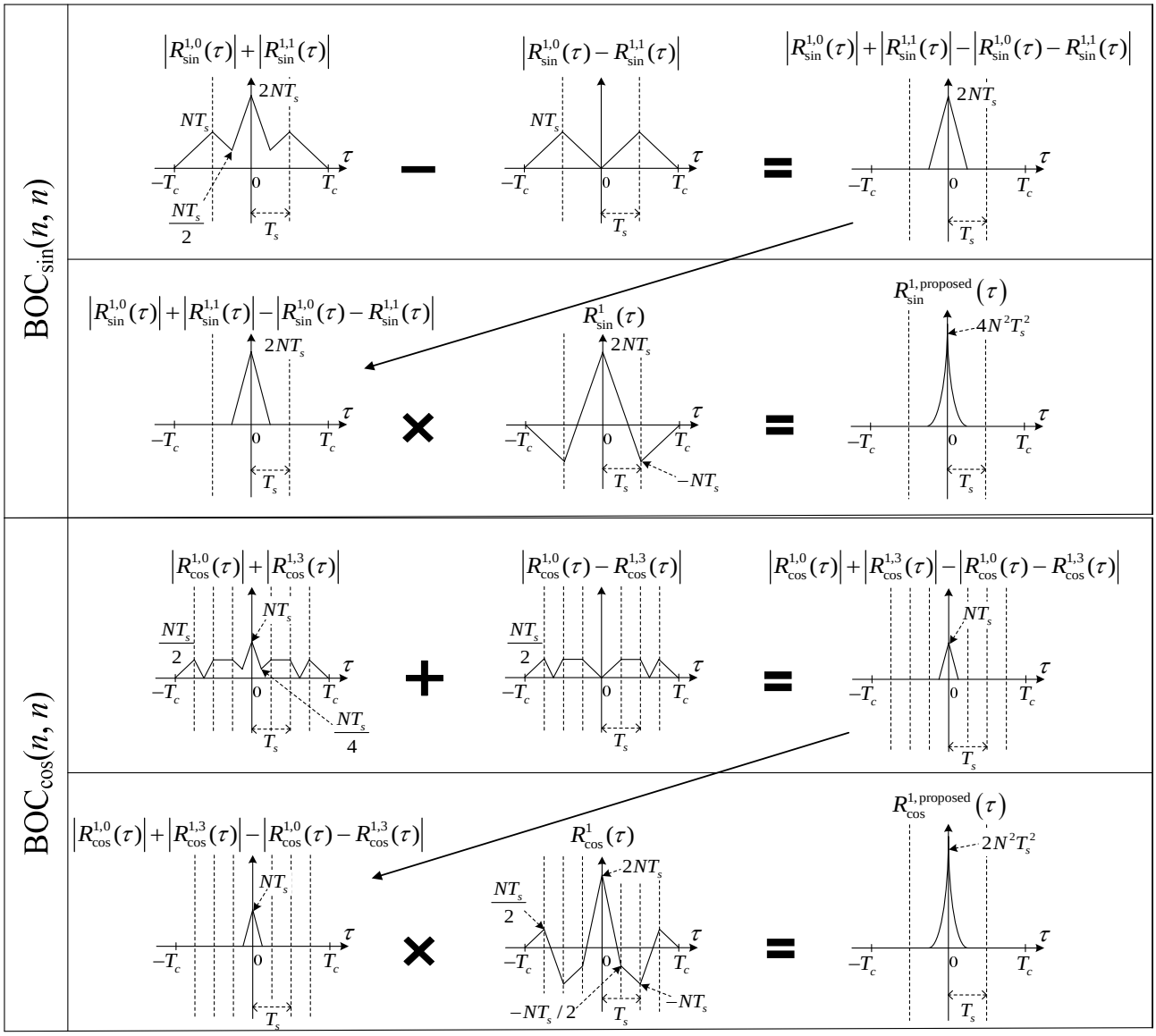


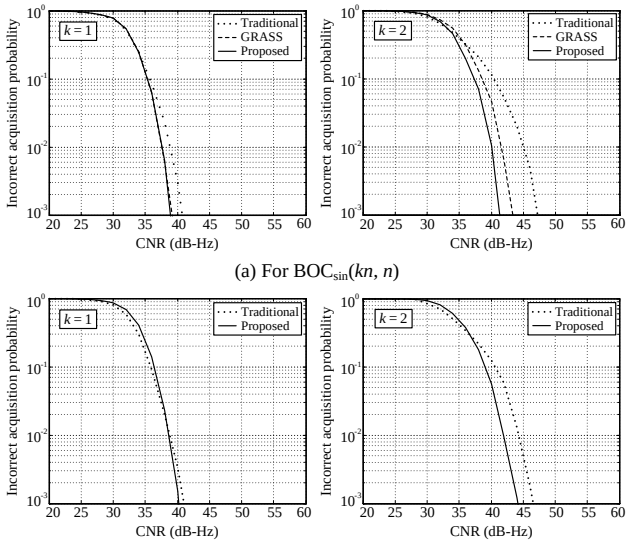
Figure 2. The umambiguous correlation functions of the proposed scheme for $BOC_{sin}(n,n)$ and $BOC_{cos}(n,n)$.

IV. NUMERICAL RESULTS

In this section, the proposed unambiguous acquisition scheme is compared with the unambiguous acquisition scheme in [16] called the general removing ambiguity via side-peak suppression (GRASS) in terms of the incorrect acquisition probability and MAT. In the comparisons, we assume the following parameters: a PRN code of $T = 127$ chips, a correlation length of $N = 127$ chips, and a search step size of T_s and $T_s/2$ for the sine-phased and cosine-phased BOC signals, respectively. For the MAT simulation, the penalty time and probability of false alarm are set to $4T$ and 10^{-3}, respectively.

Fig. 3 shows the incorrect acquisition probabilities for the proposed, GRASS, and traditional BOC schemes as a function of the carrier-to-noise ratio (CNR) for $BOC_{sin}(kn,n)$ and $BOC_{cos}(kn,n)$ when $k = 1$ and 2, where the incorrect acquisition probability is defined as the probability that any one of correlation values at $\tau \neq 0$ exceeds the main-peak magnitude of the correlation function, and the CNR is defined as P/N_0 (dB-Hz). The performance of the GRASS scheme is not shown for $BOC_{cos}(kn,n)$ since it is dedicated to the sine-phased BOC signals only. As shown in the figure, the proposed scheme yields an improvement over the GRASS scheme, and the improvement becomes larger as the value of k increases. On the other hand, the performance of the proposed scheme is slightly inferior to that of the traditional BOC scheme at relatively low CNRs (less than about 35 dB-Hz and 38 dB-Hz for

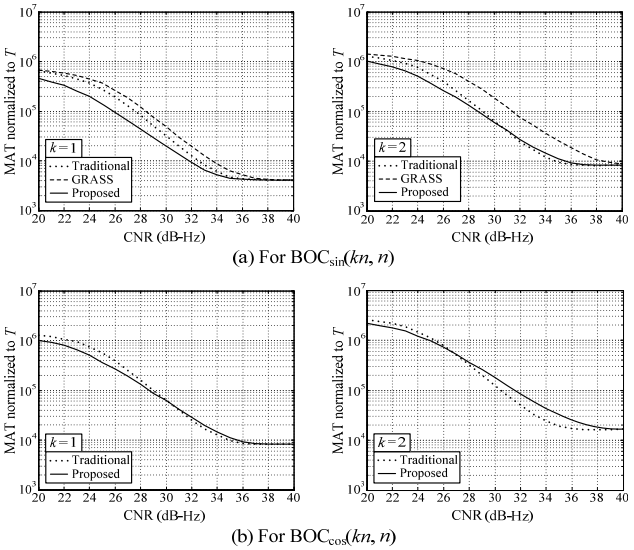
Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1



(a) For $BOC_{sin}(kn, n)$

(b) For $BOC_{cos}(kn, n)$

Figure 3. False acquisition probabilities of the proposed, GRASS, and traditional BOC schemes for $BOC_{sin}(kn,n)$ and $BOC_{cos}(kn,n)$ when k =1 and 2.



(a) For $BOC_{sin}(kn, n)$

(b) For $BOC_{cos}(kn, n)$

Figure 4. The mean acquisition time of the proposed, GRASS, and traditional BOC schemes for $BOC_{sin}(kn,n)$ and $BOC_{cos}(kn,n)$ when k =1 and 2.

$BOC_{sin}(kn,n)$ and $BOC_{cos}(kn,n)$, respectively), which can be explained as follows. A noise enhancement arises due to several absolute operations involved in the proposed scheme and its effect becomes more pronounced at relatively low CNRs, and eventually, overwhelms that of the side-peak removal of the proposed scheme, thus degrading the performance of the proposed scheme at relatively low CNRs. However, as the value of CNR increases, the side-peak removal effect would become predominant over the noise enhancement effect, thus resulting in a performance improvement of the proposed scheme over the traditional ambiguous BOC scheme. Furthermore, the improvement increases as the value of k increases.

Fig. 4 shows the MAT performances of the proposed, GRASS, and traditional BOC schemes as a function of CNR for $BOC_{sin}(kn,n)$ and $BOC_{cos}(kn,n)$ when $k = 1$ and 2. As shown in the figure, the proposed scheme outperforms the GRASS scheme in terms of MAT. Although the traditional BOC scheme has a slightly better performance than that of the proposed scheme at high CNR, the performances of the proposed and traditional BOC schemes are both good at high CNR, and thus, the small performance difference between the two schemes is insignificant at high CNR.

In addition, let us add a brief discussion on the computational complexity of the proposed and traditional schemes. From Fig. 2, we can see that the proposed scheme additionally requires three addition, one multiplication, and three absolute operations compared with the traditional scheme. It also can be seen from (10) and (11) that the sub-correlations can be obtained by collecting the correlations between the received and local BOC signals over the sub-carrier pulse duration (half duration) for $BOC_{sin}(kn,n)$ ($BOC_{cos}(kn,n)$) without any additional operation.

V. CONCLUSION

In this paper, we have proposed a novel unambiguous acquisition scheme for BOC signals in global navigation satellite systems. We have first analyzed the BOC autocorrelation function, showing the fact that it is made up of the sum of several sub-correlations shaped irregularly and which causes the multiple side-peaks of the BOC autocorrelation function. Then, we have proposed the unambiguous acquisition scheme based on a recombination of the sub-correlations. The proposed scheme is applicable to generic $BOC_{sin}(kn,n)$ and $BOC_{cos}(kn,n)$ signals, since it exploits the sub-correlations inherent in the BOC autocorrelation, regardless of the type of the BOC signal (i.e., regardless of the value of k). Finally, it has been observed that the proposed scheme removes the side-peaks completely for any sine-phased or cosine-phased BOC signal, and that it offers a performance improvement over the GRASS and traditional BOC schemes in terms of the incorrect acquisition probability and MAT.

ACKNOWLEDGMENT

This research was supported by the National Research Foundation (NRF) of Korea under Grant 2011-0018046 with funding from the Ministry of Education, Science and Technology (MEST), Korea, by the Information Technology Research Center (ITRC) program of the National IT Industry Promotion Agency under Grant NIPA-2012-H0301-12-1005 with funding from the Ministry of Knowledge Economy (MKE), Korea, and by National GNSS Research Center

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

program of Defense Acquisition Program Administration and Agency for Defense Development.

REFERENCES

[1] E. Kaplan and C. Hegarty, *Understanding GPS: Principles and Applications*, 2nd ed., Norwood: Artech House, 2006.

[2] M. Zahidul, H. Bhuiyan, E. S. Lohan, and M. Renfors, "Code tracking algorithms for mitigating multipath effects in fading channels for satellite-based positioning," *Eurasip Journal on Advances in Signal Process*., vol. 2008, article ID. 863629, 2008.

[3] J. A. Avila-Rodriguez, "On generalized signal waveforms for satellite navigation," Ph.D. dissertation, Dept. Aerospace Engineer., University of Munich, Munich, Germany, 2008.

[4] W. Liu, G. Du, X. Zhan, and C. Zhai, "MSK-binary coded symbol modulations for global navigation satellite systems," *IEICE Electron. Express*, vol. 7, no. 6, pp. 421-427, Mar. 2010.

[5] J. Wu and A. G. Dempster, "Applying a BOC-PRN discriminator to cosine phased BOC(f_s, f_c) modulation," *Electron. Lett.*, vol. 45, no. 13, pp. 689-690, June 2009.

[6] J.-C. Lin, "A modified PN code tracking loop for direct-sequence spread-spectrum communication over arbitrarily correlated multipath fading channels," *IEEE J. Selected Areas in Commun*., vol. 19, no. 12, pp. 2381-2395, Dec. 2001.

[7] J.-C. Lin, "Noncoherent sequential PN code acquisition using sliding correlation for chip-asynchronous direct-sequence spread-spectrum communications," *IEEE Trans. Commun*., vol. 50, no. 4, pp. 664-676, Apr. 2002.

[8] J.-C. Lin, "Differentially coherent PN code acquisition with full-period correlation in chip-synchronous DS/SS receivers," *IEEE Trans. Commun*., vol. 50, no. 5, pp. 698-702, May 2002.

[9] J.-C. Lin, "Differentially coherent PN code acquisition based on a matched filter for chip-asynchronous DS/SS communications," *IEEE Trans. Vehic. Technol*., vol. 51, no. 6, pp. 1596-1599, Nov. 2002.

[10] J.-C. Lin, "Low-complexity noncoherent PN code chip timing recovery with resistance to MAI for a bandlimited CDMA receiver," *IEEE Trans. Vehic. Technol*., vol. 52, no. 5, pp. 1315-1328, Sep. 2003.

[11] J.-C. Lin, "A frequency offset estimation technique based on frequency error characterization for OFDM communications on multipath fading channels," *IEEE Trans. Vehic. Technol*., vol. 56, no. 3, pp. 1209-1222, May 2007.

[12] N. Martin, V. Leblond, G. Guillotel, and V. Heiries, "BOC(x, y) signal acquisition techniques and performances," in *Proc. ION GPS/GNSS*, pp. 188-198, Portland, OR, Sep. 2003.

[13] A. Burian, E. S. Lohan, V. Lehtinen, and M. Renfors, "Complexity considerations for unambiguous acquisition of Galileo signals," in *Proc. Workshop on Positioning, Navig., and Commun.*, pp. 65-74, Hannover, Germany, Mar. 2006.

[14] E. S. Lohan, A. Burian, and M. Renfors, "Low-complexity unambiguous acquisition methods for BOC-modulated CDMA signals," *Int. J. Sate. Commun. Networking*, vol. 26, no. 6, pp. 503-522, Nov.-Dec. 2008.

[15] O. Julien, C. Macabiau, M. E. Cannon, and G. Lachapelle, "ASPeCT: unambiguous sine-BOC(n,n) acquisition/tracking technique for navigation applications," *IEEE Trans. Aerospace and Electron. Syst*., vol. 43, no. 1, pp. 150-162, Jan. 2007.

[16] Z. Yao, M. Lu, and Z. Feng, "Unambiguous sine-phased binary offset carrier modulated signal acquisition technique," *IEEE Trans. Wireless Commun*., vol. 9, no. 2, pp. 577-580, Feb. 2010.

[17] Y. Lee, J. Baek, and S. Yoon, "Cyclostationarity 를 갖는 부상관함수들의 재조합에 기반을 둔 BOC 신호 획득 기법," *J. Korea Inform. Commun. Society*, vol. 36, no. 7, pp. 459-466, July 2011.

[18] F. D. Nunes, M. G. Sousa, and J. M. N. Leitao, "Gating functions for multipath mitigation in GNSS BOC signals," *IEEE Trans. Aerospace and Electron. Syst*., vol. 43, no. 3, pp. 951-964, July 2007.

[19] E. S. Lohan, A. Lakhzouri, and M. Renfors, "Feedforward delay estimators in adverse multipath propagation for Galileo and modernized GPS signals," *Eurasip Journal on Applied Signal Process*., vol. 2006, article ID. 50971, 2006.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Multicast Routing in Wireless Mesh Networks

Jakub Sobczak
Faculty of Electronics and Telecommunications
Poznan University of Technology
Poznan, Poland
e-mail: jakub.sobczak@doctorate.put.poznan.pl

Piotr Zwierzykowski
Faculty of Electronics and Telecommunications
Poznan University of Technology
Poznan, Poland
e-mail: piotr.zwierzykowski@put.poznan.pl

***Abstract*—Wireless networks have recently gained on significance. In environments where it is impossible to build traditional infrastructure wireless networks, ad hoc and wireless mesh networks are used. The purpose of this paper is to evaluate certain multicast routing algorithms used in ad hoc and wireless mesh networks. The first part of the paper addresses the subject of ad hoc and wireless mesh networks, as well as the issue of multicasting in these networks. Furthermore, the paper contains a review of routing algorithms. In the last part of the paper the conducted research is presented and a multicast algorithms with the best performance in WMNs is chosen.**

Keywords-WMN; multicast routing algorithms

I. INTRODUCTION

The growing need for unlimited Internet access and the constant progress in terms of developing new technologies (e.g., smartphones, tablets) caused a considerable development of wireless computer networks. Thanks to this method, access to the Internet has become less expensive, which, in turn, has been followed by a rapid growth in the number of people using it.

Until recently, there have been two basic approaches to creating wireless computer networks: infrastructure and ad hoc. Infrastructure mode requires the use of wireless access points that act as a go-between in conveying information and provide control over the process of communication. In contrast, wireless network operating in ad hoc mode does not require any superior or control devices - every device connected to this network may fulfill the same functions. However, in ad hoc networks, all devices have limited resources - not only energy, but also bandwidth.

The purpose of this research is to examine and compare algorithms and protocols being used in Wireless Mesh Networks in homogeneous conditions and the same parameters. The motivation for this study is the lack of such a comparison in subject literature.

The paper is organized as follows. Section II describes wireless mesh networks (WMN). The next section presents multicast algorithms and algorithms used in WMN. The following section shows simulation parameters whereas the results are discussed in Section V. Finally, Section VI concludes the paper.

II. WIRELESS MESH NETWORKS

There are three main types of ad hoc networks (Fig. 1).

A. Mobile ad hoc Networks (MANETs)

One of the most popular applications of ad hoc networks is using them as mobile networks. Mobile ad hoc network is created dynamically by a group of mobile devices without any assistance of the existing infrastructure. In such a network, devices communicate between one another by pursuing one or more hops (Fig. 2).

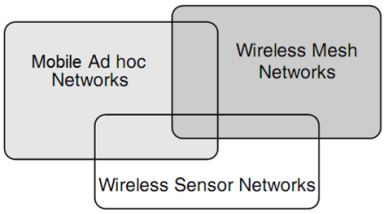


Figure 1. Types of ad hoc networks

Main advantages of MANETs are their flexibility deriving from their dynamic structure and independence of any fixed infrastructure. Unfortunately, it comes with a price, because mobility of devices connected to such a network influences the way of designing routing algorithms and algorithms for these networks - mainly because such networks are less stable and prone to disconnect. In the case of proactive algorithms, the status of the network has to be refreshed quickly enough to keep up with changes in the structure of the network. If this requirement is not met, the packet loss may increase on the one hand, but on the other, refreshing topology information too often may influence the links load and reduce effectiveness of network resources usage. Thus, on-demand (reactive) routing algorithms are much more efficient and give better results in such networks.

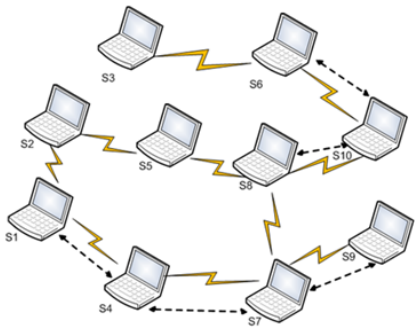


Figure 2. Example of MANET

B. Wireless Sensor Networks - WSNs

A network that consists of spatially distributed autonomous sensors that perform common tasks is called a *sensor network*. Usually, these types of networks are used to monitor environmental conditions, such as temperature, humidity, pressure, etc. Every single sensor has the ability

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

to process data and send them wirelessly to other sensors in range. The development of WSNs was motivated by military applications - for intelligence, battlefield monitoring, etc. Nowadays they are also used in consumer applications, such as security monitoring, monitoring weather conditions or even traffic.

C. Wireless Mesh Networks -WMNs

Wireless Mesh Network is a very specific ad hoc network, which consists of two basic elements: *mesh backbone* and *customers* (Fig. 3). WMNs are the most static ad hoc networks when it comes to topology and structure. The backbone is created between wireless, but static, *mesh routers* (MRs), which have neither bandwidth nor energy limits. Some of MRs, which have a cable connection to the Internet, are called *Internet gateways* (IGWs) - which resembles access points in a traditional infrastructure mode. In addition, reliability of the network is improved by transmitting data between nodes in a *mesh* way, which, when combined with being independent from the local infrastructure, makes WMNs perfect to be applied in places where building a traditional cable network infrastructure would be too expensive or impossible (e.g., desert).

Due to the fact that nodes creating a backbone of the network are mostly stationary, routing methods that establish a permanent connection between multiple nodes, may be applied. In the majority of cases, there is no direct connection between each node in WMN, but they are able to communicate using neighboring nodes. What is more, WMNs have the ability of self-configuration and repair in case the position of a node changed or nodes were added or removed.

As WMNs, similarly to ad hoc networks, do not depend on the available telecommunications infrastructure, they are a very good choice whenever a decentralization is required. Moreover, they may be used wherever it is necessary to quickly restore communication, e.g., in natural disaster areas, where telecommunications infrastructure has been damaged or destroyed.

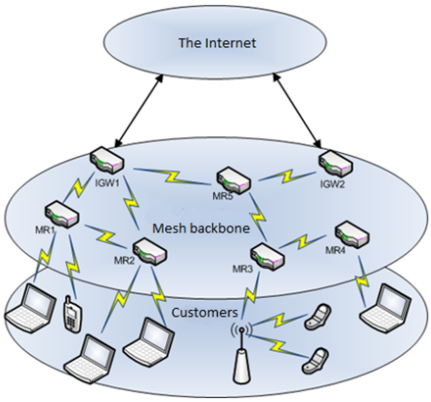


Figure 3. Wireless Mesh Network architecture

III. MULTICAST ROUTING

A. Types of transmission in packet networks

In packet networks using Internet Protocol we can distinguish four main types of transmission as far as the way of delivering information is concerned: unicast, broadcast, anycast, and multicast.

Multicast transmission is based on sending the same data stream to multiple receivers. Its main advantage is that even though the information is sent to a group of receivers, it is transmitted by each network link only once, which saves considerable amount of bandwidth and energy and eliminates the need for sending multiple copies of the data. To be able to use multicast in WMNs it is necessary to implement algorithms generating a structure of the network as well as ones choosing optimal routes between devices.

In subject literature, 'algorithm' and 'protocol' definitions are often used in an ambiguous way. In this article, authors assume that only such a solution should be referred to as 'protocol', for which a least state machine is defined.

B. Multicast routing in MANETs

Multicast routing algorithms are divided into the three following groups: *mesh-based*, *tree-based* and *hybrid*.

Mesh-based algorithms are recognized as the most reliable, because they create a structure in which more than only one path can connect the sender with the receiver. In tree-based algorithms only one path from sender to receiver exists, but it makes the routing much more effective and eliminates the possibility of loops occurring in the network. Moreover, both - the sender (*sender-initiated*) and the receiver (*receiver-initiated*) may initiate creation of the *multicast tree*.

Typical *tree-based* algorithms are: MAODV [1] and AMRIS [2], whereas typical *mesh-based* algorithms are: ODMPR [3] and CAMP [4]. The existing research [5] shows that in MANET environments, where changes in topology are common, *mesh-based* algorithms show better results than *tree-based* algorithms, which is due to the existence of redundant links in the mesh structure.

As this paper presents only the initial stage of this research, only mesh- and tree-based protocols are analyzed.

C. Multicast routing in WMNs

Wireless Mesh Networks are a relatively new wireless technology and that is why the available literature does not show any recent research studies that would compare the existing multicast routing algorithms. One of the first studies on the topic is [8], in which Ruiz states that the *shortest path tree* (SPT) algorithm does not work well for WMNs and proposes minimum number of transitions (MNT) algorithm that focuses on using properties of multicast transmission to reduce the number of transmissions necessary to reach all nodes in a tree.

In [6], Nguyen and Xu present their analysis on effectiveness of *Minimal Coverability Tree* (MCT) and SPT algorithms. The research shows that SPT algorithms are much more efficient than MCT algorithms.

Moreover, two multicast algorithms for WMNs, i.e., *Level Channel Assignment* (LCA) and *Multi-Channel Multicast* (MCM) are introduced in [7]. These algorithms aim not only to increase throughput in WMNs, but also to minimize the number of hops in a tree. Multicast *mesh tree* is created by dividing routers to different levels using *Breadth First Search* (BFS) algorithm as well as heuristic channel allocation to different radio interfaces.

Zhao, et al. [8] proposes *Gateway Associated Multicast Protocol* (GAMP), which was created to improve *Quality of Service (QoS)* in Wireless Mesh Networks. GAMP is

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

a hybrid algorithm, because the sender broadcasts *Hello* messages to all active access nodes and when a receiver wants to join a group, it sends a connection message to the access node (*on-demand*).

D. Algorithms chosen for the simulation

Several factors were taken into consideration while choosing specific multicast routing algorithms for the research: clarity and granularity of the description of an algorithm, comparability to other algorithms and complexity of the implementation. Basing on the aforementioned criteria, the following algorithms were chosen: MAODV (MANETs), ODMPR (MANETs), MNT (WMNs), MCM (WMNs), LCA (WMNs). However, the LCA algorithm was omitted because its comparison with the MCM algorithm is available [7] and shows that the MCM algorithm gives better results than LCA.

E. Description of the algorithms and protocols

Multicast Ad hoc On-Demand Distance Vector (MAODV)
MAODV [1] is a reactive (on-demand) *tree-based* algorithm. It enables fully dynamic and multi-hop routing between mobile nodes willing to join a multicast group in ad hoc networks. What makes this algorithm different from the other ad hoc multicast routing algorithms is that each multicast group has its own sequence number assigned by a group leader (root of the tree). This number increases in time, what guarantees choosing always the most up-to-date paths, because nodes choose paths with the highest sequence. What is more, the group leader sends *Group Hello* messages to all members of the group to update its status.

Because MAODV is a reactive (on-demand) algorithm, as long as the connection between members of a multicast group is preserved, no actions are taken. Each node monitors the state of next-hop links, which is why in the case any path fails, it may be quickly restored.

On-Demand Multicast Routing Protocol (ODMRP)
ODMRP [3] is an *mesh-based* algorithm which implements forwarding group concept for multicast routing. It means that only some nodes of the multicast group may forward and transmit packets. The topology of the whole network is never stored anywhere, which means that user management is dynamic and *on-demand*. For managing routing activities, it requires storing some data structures, such as: *routing table*, *forwarding group table*, *group members table* and *message cache* in different types of nodes.

To keep the mesh structure in the most up-to-date state, soft state approach is used. This means that in the case a source wants to leave the multicast group, it just stops broadcasting *Join Query* messages and, if a receiver wants to leave a group, it stops broadcasting *Join Reply* messages. After some time, a timeout occurs.

Minimum number of transmissions (MNT)
MNT algorithm is described in [9]. According to Ruiz, the general assumption that Steiners tree is a tree of minimal cost, is not necessarily true in WMNs. Ruiz redefines the issue of multicast tree minimization in ad hoc networks to reduce the amount of data transmission. Existing calculations explicitly assume that, given node v, it is necessary to send multicast data k-times to reach k-neighbors of the node v. However, using multicast transmission, it is enough to send multicast transmission *only once* to reach any number of nodes v neighbors. Thus, the minimal cost tree is *not* the one which stores the lowest cost of each edge, but the one connecting senders and receivers in the least number of transmissions needed. This type of structure has been defined by Ruiz as *minimal data overhead tree*.

Multi-Channel Multicast (MCM)
MCM first builds a multicast structure by minimizing the number of relay nodes and hop count distances between the source and destinations, and then uses dedicated channel assignment strategies to improve the network capacity by reducing interference [7]. The authors of the algorithm have made an observation that when all the nodes have multiple radio interfaces, the multicast problem becomes, in fact, a special case of broadcast.

The first step of MCM is realized by *breadth first search* (BFS) algorithm. Then, all edges between any two nodes of the same level are deleted and a *tree mesh* is built.

In the second step of the algorithm, the minimal number of relay nodes forming a broadcast tree is identified. In the tree mesh one node may have more than one parent. The purpose of this step is to identify the only parent for each node, so that the number of relay nodes stays minimal.

After creating a multicast structure thanks to which each multi-receiver may connect with the gateway through minimal hop count distance, the algorithm assigns channels to the interfaces of the tree nodes using two allocation algorithms: *ascending channel allocation* and *heuristic channel allocation* [7].

IV. Simulation parameters

The quality of the simulation is directly related to a simulation model. In the case of WMNs, this model is complicated and consists of five sub-models:

- **node** - defining its hardware and software,
- **arrangement and mobility** (of the topology) - it ensures proper arrangement of nodes,
- **radio** - describing characteristics of the radio interface of a node,
- **propagational** - describing attenuation and radio channel characteristics,
- **traffic** - defining traffic in the network.

Some of the sub-models are based on actual measurements, e.g., propagational and traffic. The rest is synthetic and arbitrary, like the topology generator.

A. Network topology generator

In this study, as a topology generator, we use the algorithm called *Node Placement Algorithm for Realistic Topologies* (NPART) [10] to preserve reality of the generated topology.

The NPART algorithm was created on the basis of the measurements conducted on actual active WMNs in Berlin and Leipzig. The authors of NPART proposed this algorithm because they made an observation that it is difficult to find an algorithm with output similar to real networks.

B. Network topology used in simulations

It is assumed that nodes are allocated on a plane 1000 x 1000 units and that the communication radius is 100. The cost of each connection varies between 10 and 100 and the delay metrics is the Euclidean distance between nodes. For the simulation to be as realistic as possible, 1000 topologies have been generated.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

C. Parameters of the simulation

Node count. The total number of nodes in the network is a very important parameter in the process of building structures by multicast routing algorithms. In the study, networks of 50, 100, 150 and 200 nodes were analyzed.

The number of multicast groups. In real WMNs, any number of multicast groups may exist. However, to simplify the analysis of the results, only one multicast group will exist in a simulated network.

The size of a multicast group. As the multicast group grows, finding the optimal structure becomes more time-consuming and requires more hardware resources. Groups of 5, 10, 15 and 20 nodes were examined in the research.

D. Parameters of multicast routing examined in the research study

The mean path length between the sender and a multicast group. Multicast routing algorithms create a structure that enables the most efficient transmission between sending and receiving nodes in WMNs. It is expected that the paths will be as short as possible, which means as few relaying nodes as possible. Each relaying node increases the risk of path breakdown and introduces additional delay and cost. In this research, the mean path length parameter is calculated by adding up all path lengths between sender and each separate node of a multicast group and dividing the result by the number of nodes in the particular multicast group. The path length is expressed in the number of edges (NE) between the sender and the receiver.

The mean path cost between the sender and a multicast group. The value of this attribute reflects the whole set of parameters describing a cost of creating an edge between two nodes. As such a parameter we could assume, for example, bandwidth required for the transmission to be successful. In simulations conducted in this research, the cost of each connection varies between 10 and 100 and is chosen in a random way. Thus, the mean path cost parameter is calculated by adding up all path costs between sender and each separate node of a multicast group and dividing the result by the number of nodes in the particular multicast group. The cost of the path is expressed in Cost Unit (CU).

The mean path delay between the sender and a multicast group. It is desirable for delays in a transmission between the sender and the receiver to be minimal in most cases of modern multicast connections applications. Delay is a time necessary to transmit data from one node to the other. In this paper, delay metrics is assumed to be the Euclidean distance between two nodes and is expressed in Delay Units (DU). Thus, the mean path delay parameter is calculated by adding up all path delays between sender and each separate node of a multicast group and dividing the result by the number of nodes in the particular multicast group.

V. RESULTS

A. The influence of network size on the performance of multicast algorithms

The influence of the number of nodes on the performance of multicast algorithms was examined first. The results are shown below – Figs. 4-6.

The results clearly show that MNT algorithm stands out as compared to other algorithms. This is due to the specific way the algorithm decides to create a path. MNT chooses nodes which cover as many receivers as possible and takes into consideration the rest of the parameters afterwards.

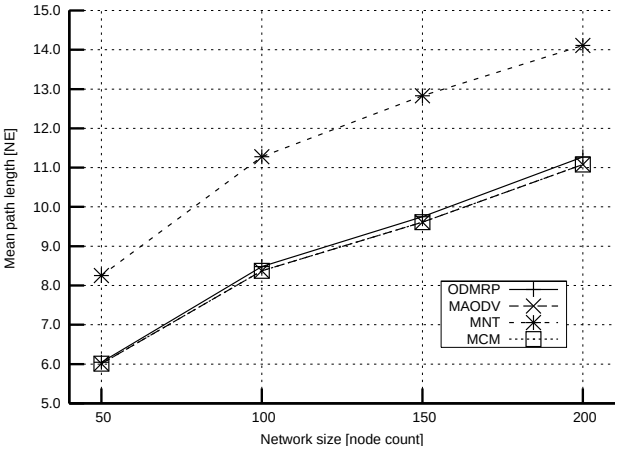


Figure 4. Mean path length in function of network size

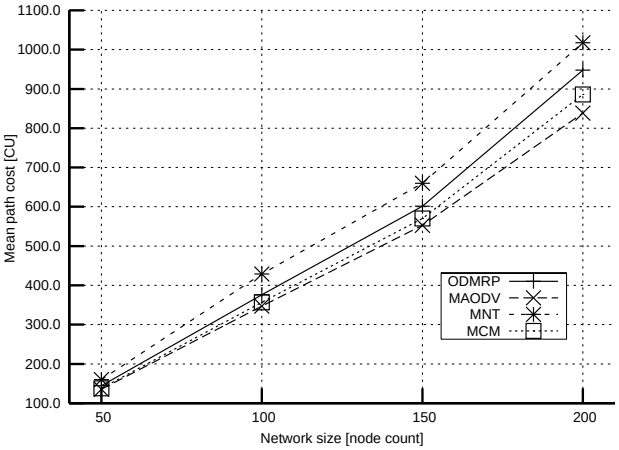


Figure 5. Mean path cost in function of network size

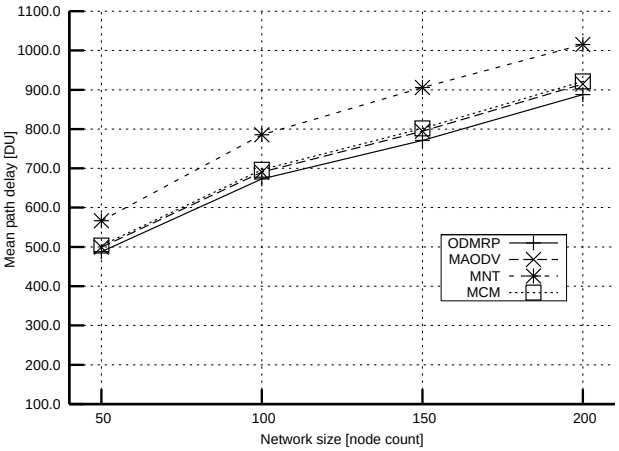


Figure 6. Mean path delay in function of network size

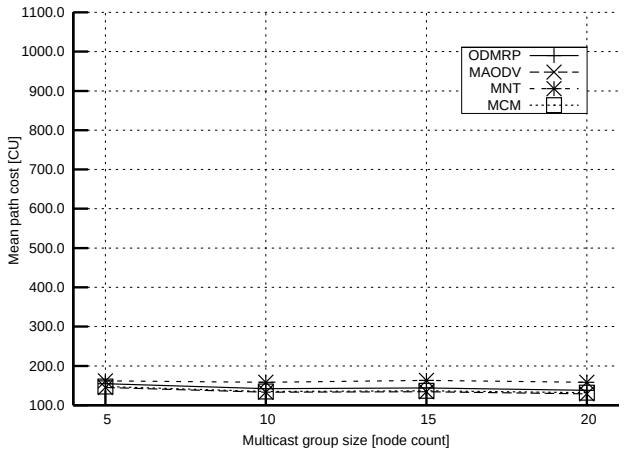


Figure 7. The influence of the number of receiving nodes on the mean path cost (50 nodes)

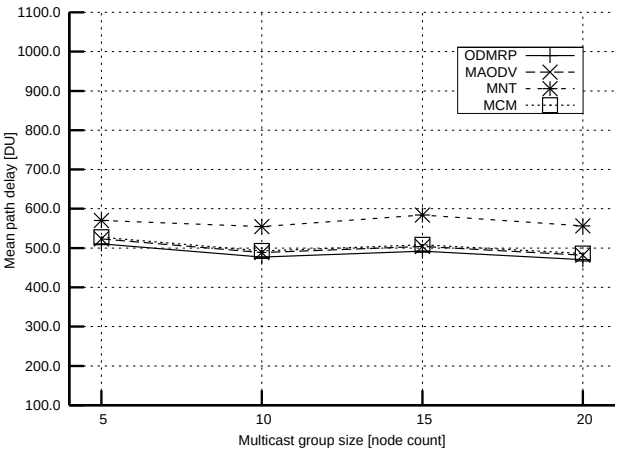


Figure 10. The influence of the number of receiving nodes on the mean path delay (50 nodes)

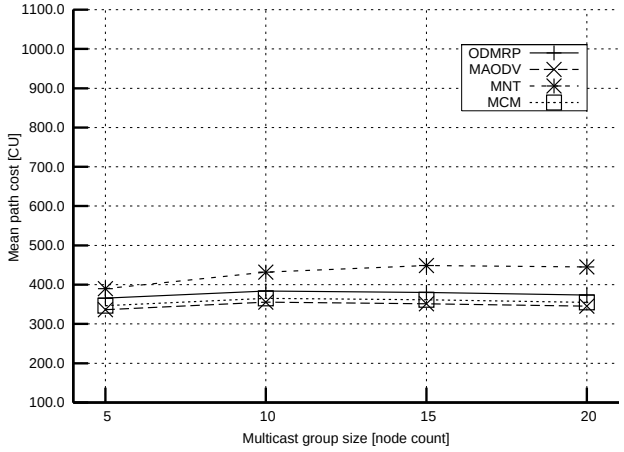


Figure 8. The influence of the number of receiving nodes on the mean path cost (100 nodes)

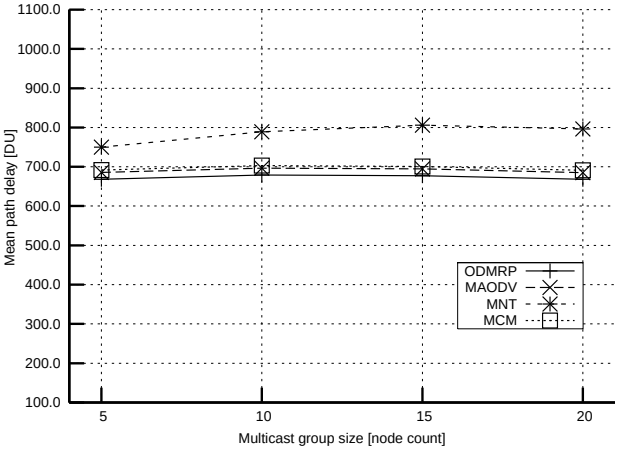


Figure 11. The influence of the number of receiving nodes on the mean path delay (100 nodes)

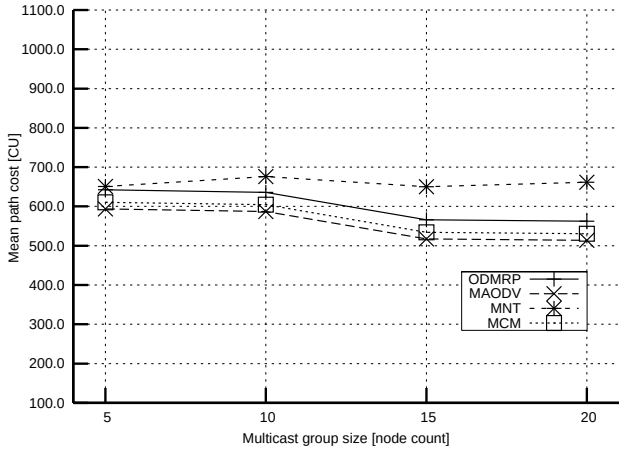


Figure 9. The influence of the number of receiving nodes on the mean path cost (150 nodes)

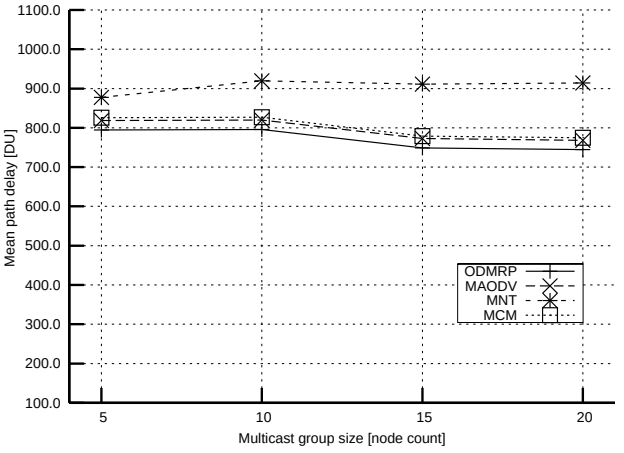


Figure 12. The influence of the number of receiving nodes on the mean path delay (150 nodes)

Surprisingly, despite the fact that ODMRP, MAODV and MCM have different criteria of choosing nodes to join a path, *mean length of the path* is very similar in each case.

This is probably because of the way the cost of the path

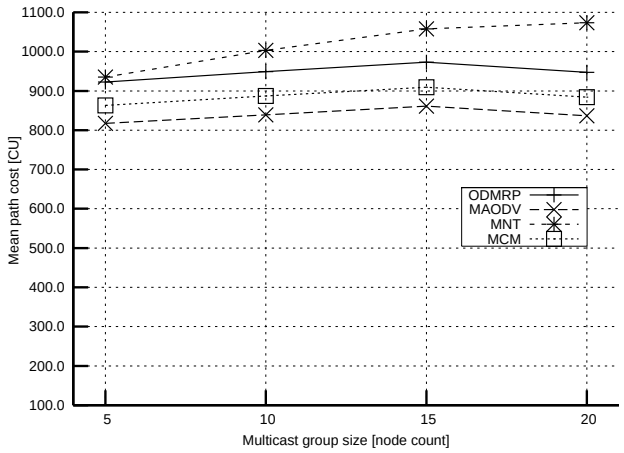


Figure 13. The influence of the number of receiving nodes on the mean path cost (200 nodes)

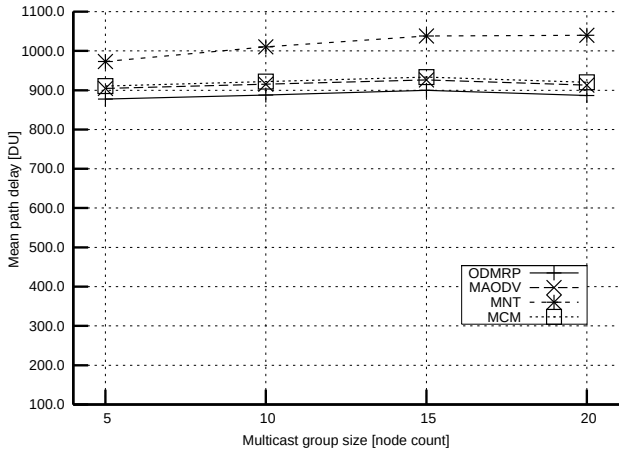


Figure 14. The influence of the number of receiving nodes on the mean path delay (200 nodes)

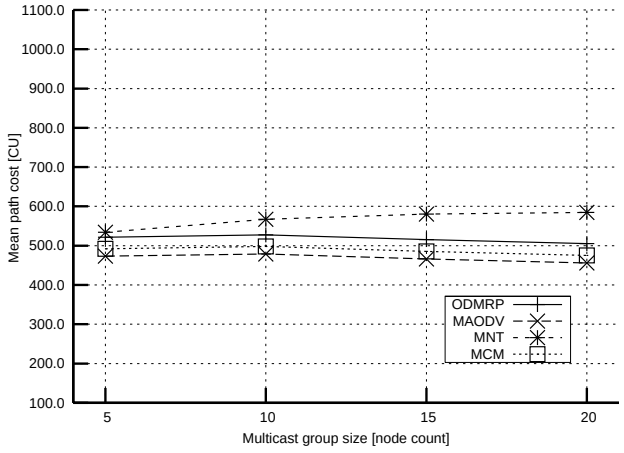


Figure 15. The influence of the number of receiving nodes on the mean path cost (averaged)

is defined. A very similar situation occurs when *mean path delay* is considered – the algorithm which achieved the best

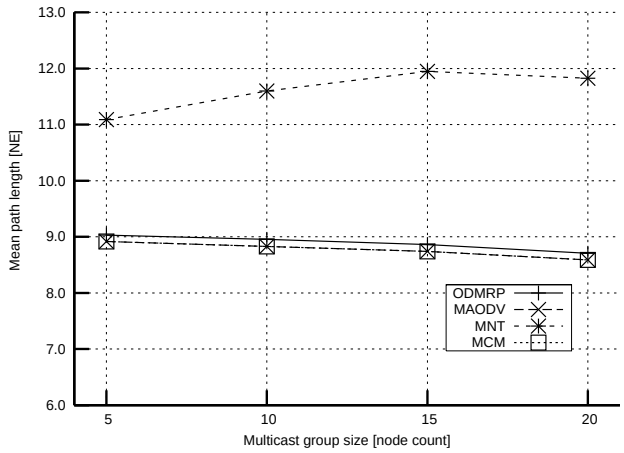


Figure 16. The influence of the number of receiving nodes on the mean path length (averaged)

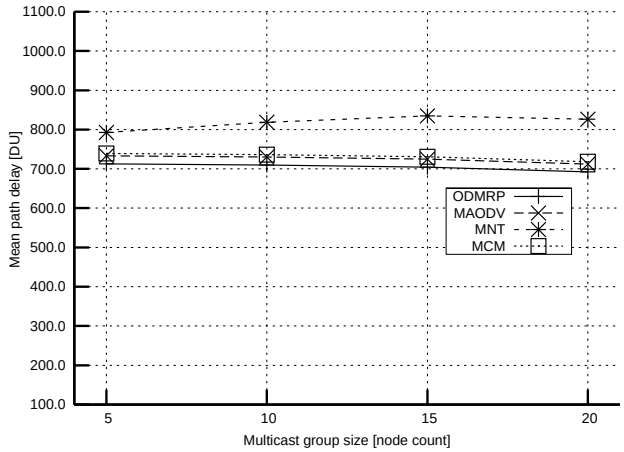


Figure 17. The influence of the number of receiving nodes on the mean path delay (averaged)

results was ODMRP, because this algorithm takes time into consideration while creating the structure.

Concluding, although there are differences between the algorithms, in some cases ODMRP, MAODV and MCM performance is similar. Considering all of the parameters examined, the MAODV algorithm performed slightly better than the others, whereas MNT proved to be the least effective.

B. The influence of the multicast group size on performance of multicast algorithms

The study of the influence of the multicast group size on the performance of multicast algorithms was performed for 50, 100, 150 and 200 nodes and for group sizes of: 5, 10, 15 and 20 (Figs. 7-14).

Network consisting of 50 nodes. Figs. 7 and 10 show that the mean path cost is similar for each algorithm. Together with the growth of the receiving group, the cost drops slightly, which may be caused by the fact that shorter paths to nodes closer to the source might have appeared. Only in the case of the MNT algorithm, the path cost decrease is less dynamic, but this is caused by the algorithm of path

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

construction. The increase in path delay, cost and length for 15 receiving nodes is not surprising, because the values should fluctuate within certain range.

Network consisting of 100 nodes. In the network consisting of 100 nodes the mean path cost and delay change in a different way than in a smaller network. Figs. 8 and 11 show the values of these parameters that seem to stay within a certain range for all of the algorithms except for MNT. As suspected, MNT behaved in a different way - values of the mean cost, delay and path length increase almost steadily.

Network consisting of 150 nodes. In the case of a network consisting of 150 nodes, an improvement may be observed as compared to a network consisting of 100 nodes. MAODV, ODMRP and MCM algorithms show a significant drop in the mean path delay and cost (Figs. 9 and 12). Even though it may indicate that these algorithms are very effective in large networks with an increasing number of receivers, it might also mean that members of multicast groups were chosen in an unfavorable way.

Network consisting of 200 nodes. The results of the evaluation for the network consisting of 200 nodes confirm the previous findings and are presented in Figs. 13 and 14.

C. Conclusions

Despite the differences between ODMRP (*mesh-type*), MAODV (*tree-type*) and MCM (*tree-type*) algorithms their performance is comparable. However, it should be pointed out that as the best candidate to be used in multicast routing the **MAODV** algorithm should be chosen. The analysis and research conducted in this paper show that this algorithm, from among all the tested algorithms, offers the lowest mean path cost (Fig. 15) and the lowest mean path lengths (Fig. 16), only performing slightly weaker than ODMPR as when comes to the lowest mean path delay (Fig. 17).

VI. Conclusions and future works

The paper presents performance evaluation of the selected multicast routing algorithms for WMNs. The most interesting results were obtained for MAODV and MNT algorithms.

MAODV uses a *tree* structure to operate, which makes its efficiency in WMNs surprising, because the nature of WMNs suggests that *mesh-based* algorithms should perform better in these networks. Moreover, MAODV was not designed for WMNs, but for ad hoc networks. This makes MNT algorithms results even more peculiar, because it is a *mesh-based* algorithm specifically designed for a multihop environment. However, the analysis of the mechanisms used by this algorithm suggests that it is not possible for this algorithm to achieve results even remotely comparable to other tested algorithms.

It is worth mentioning that the MAODV algorithm is used as an example of a multicast routing algorithm by the IEEE 802.11s workgroup responsible for standardizing *Wireless Mesh Networks*. This paper presents the initial stage of the research during which the authors evaluated and compared discussed in literature protocols and algorithms for WMNs. Considering the fact that this paper analyzes only mesh- and tree-based protocols, further papers on the topic of hybrid protocols in WMNs shall follow. In the next stage of this research, the most efficient algorithms will be compared to the solutions proposed by the authors.

References

[1] E. M. Royer and C. E. Perkins, *Multicast operation of the ad hoc on-demand distance vector routing protocol*, Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking (MobiCom '99), ACM, New York, USA, pages 207-218, 1999.

[2] C. W. Wu and Y. C. Tay, *AMRIS: A Multicast Protocol for Ad Hoc Wireless Networks*, Proceedings of the IEEE Military Communications Conference (MILCOM), Atlantic City, USA, pages 25-29, 1999.

[3] S. J. Lee, W. Su, and M. Gerla, *On-Demand Multicast Routing Protocol in multihop wireless mobile networks*, ACM/Kluwer Mobile Networks and Applications vol. 7, no. 6, pages 441-453, 2002.

[4] J. J. Garcia-Luna-Aceves and E. L. Madruga, *The Core-Assisted Mesh Protocol*, IEEE Journal on Selected Areas in Communications, Volume 17, Issue 8, pages 1380-1394, 1999.

[5] S. J. Lee, W. Su, J. Hsu, M. Gerla and R. Bagrodia, *A Performance Comparison Study of Ad Hoc Wireless Multicast algorithms*, Proceedings of IEEE INFOCOM, vol. 2, pages 565-574, 2000.

[6] U. T. Nguyen, and J. Xu, *On multicast routing in wireless mesh networks*, Computer Communications, Volume 31, Issue 7, pages 1385-1399, 2008.

[7] G. Zeng, B. Wang, Y. Ding, L. Xiao and M. Mutka, *Efficient Multicast Algorithms for Multichannel Wireless Mesh Networks*, IEEE Trans. Parallel Distrib. Syst., vol. 21, no. 1, pages 86-99, 2010.

[8] L. Zhao, A. Al-Dubai, and X. Liu, *A new multicast routing algorithm for the Wireless Mesh Networks*, 9th Annual Postgraduate Symposium on Convergence of Telecommunications, Networking and Broadcasting, Liverpool, pages 151-156, 2008.

[9] P. M. Ruiz, and A. F. Gomez-Skarmeta, *Approximating Optimal Multicast Trees in Wireless Multihop Networks*, Proceedings of the 10th IEEE Symposium on Computers and Communications, La Manga del Mar Menor, Spain, pages 686-691, 2005.

[10] B. Milic, and M. Malek, *NPART - Node Placement Algorithm for Realistic Topologies in Wireless Multihop Network Simulation*, Proceedings of the Second International Conference on Simulation Tools and Techniques (Simutools '09), Institute for Computer Sciences, Brussels, Belgium, pages 9.1-9.10, 2009.

[11] G. Agacinski, *Multicast Routing in Wireless Mesh Networks*, Poznan University of Technology, 2011.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Direction-based Greedy Forwarding in Mobile Wireless Sensor Networks

Riad Kouah[1,2], Samira Moussaoui[2], and Mohamed Aissani[1]
[1] Research Unit in Computer Science, Ecole Militaire Polytechnique (EMP)
P.O. Box 17 EMP, Bordj-El-Bahri, 16111, Algiers, Algeria
riadkouah@hotmail.com, maissani@gmail.com
[2] Computer Science Department, University of Science and Technology (USTHB)
P.O. Box 32 USTHB, El-Alia, Bab Ezzouar, Algiers, Algeria
rkouah@usthb.dz, smoussaoui@usthb.dz

Abstract— **Geographical routing in mobile wireless sensor networks has attracted big attention in recent years by introducing new challenges. When a node has a packet to forward, it selects the closest available neighbor to the sink as the next forwarder regarding only the location parameter. However, this routing strategy does not consider the topology changes caused by the mobility of nodes, which may degrade performance or cause failures. To overcome this problem, we propose an efficient greedy forwarding mechanism based on a new decision metric that considers the distance to the sink, the moving direction and the moving speed of the forwarding candidate neighbors of a sender node. The moving direction depends on both distance and angle of a neighbor according to the sink between two successive location beacons. Associated with the well-known GPSR routing protocol, our proposal achieved good performance in terms of packet delivery ratio, average path length, control packet overhead and energy consumption.**

Keywords—Mobile sensor networks; geographical routing; node mobility; greedy forwarding.

I. INTRODUCTION

Currently, Wireless Sensor Networks (WSNs) attract the attention of many researchers due to the various challenges imposed by sensor nodes, such as the small amount of available memory, the limited processing capability, the limited lifetime of batteries and the small range of radio transceiver. Moreover, several researches are focusing on mobile WSNs where node mobility is critical to meet applications requirements.

Today, by introducing mobility to WSNs, we can further improve the network capability in many aspects [1, 2]. Since the deployment of WSNs had never been considered completely static, the node mobility problem imposes various challenges to deal with, such as the connectivity, the coverage, the energy consumption and the routing. The later challenge has involved the design of several new protocols, especially geographical routing protocols [3-5]. However, only few of these protocols are designed for mobile WSNs. In fact, the efficient and scalable greedy forwarding is a promising scheme for large-scale WSNs when node locations are available [6]. Indeed, the packet is forwarded to a 1-hop neighbor who is closer to the destination than the actual node. This process is repeated until the packet reaches its destination.

Traditionally, the selection of the next forwarder node is based only on the location parameter assuming ideal link conditions. However, this can degrade the performance of geographical routing in mobile WSNs. The location failure, which results from nodes' mobility, degrades the routing performance or may lead to forwarding failures. Indeed, when a node selects its forwarder node, based only on location information, there may be another neighbor having better conditions according to the sink in terms of location, moving direction and moving speed. This effect is accentuated when the selected forwarder is within the range limit of the sender node, which probably leads to a forwarding failure. In this case it is better to select as forwarder the neighbor ensuring the tradeoff between location, moving direction and moving speed.

To overcome the limits of the existing greedy forwarding schemes proposed for mobile WSNs, we first analyze the impact of node mobility on the routing performance. Then we propose an efficient greedy forwarding mechanism, called DGF (Direction-based Greedy Forwarding), which combines the location information, the moving direction and the moving speed of the forwarding candidate neighbors when a sender node selects the next forwarder of the current packet. We associate the DGF mechanism with the GPSR [3] routing protocol to use in mobile WSNs.

The rest of this paper is organized as follows. Section II presents some geographical routing schemes proposed in the literature for mobile WSNs. Section III discusses the node mobility effect on the greedy forwarding performance. Section IV describes the proposed DGF mechanism for mobile WSNs. Section V evaluates performance of our proposal. Section VI concludes the paper.

II. RELATED WORK

Node mobility arises additional challenges in WSNs. It has to be handled even in quasi-static WSN where few nodes may be mobile. One of these challenges is routing. We distinguish three classes of routing protocols in mobile WSNs based on the type of nodes: 1) protocols for static sensors and mobile sink(s) [7-9], 2) protocols for mobile sensors and static sink(s) [10-14], and 3) protocols for mobile sensors and mobile sink(s) [15]. In literature, the

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

majority of research works have been focused on the first class of protocols, while less works dealt with both the second and the third class.

Luo et al. [7] propose the two-tier data dissemination (TTDD) protocol, which is used to forward a packet from static sensors towards a mobile sink. In TTDD, sinks are assumed to be mobile with unknown and uncontrolled mobility. The data about each event are assumed to originate from a single source. Each active source creates a grid structure dissemination network over the static network, with grid points acting as dissemination nodes. A mobile sink, when it issues queries for information, it sends out a locally controlled flood to discover its nearest dissemination point. Then the query is routed to the source node through the overlay network by suing the dissemination point.

Fodor et al. [8] propose a gradient-based routing protocol (GBRP) to use mobile sinks that move in order to decrease the energy consumption of the whole network. In GBRP, sensor nodes maintain a list of neighboring next hops that are in the right direction towards the closest sink. The protocol uses a restricted flooding to update the locations of the mobile sinks. The principle behind this is to register by each node the cost between the appropriate sink and the given node and to update these routing entities only when the relative change is above a threshold.

Wang et al. [9] propose a mobile sink cluster-based routing protocol (MSRP) for WSNs. The protocol operates in four phases: clustering, registering, dissemination and maintenance. The network is divided into multiple clusters during the first phase. The mobile sink which comes into the communication range of a cluster-head is registered into this cluster using the second phase. Once the mobile sink is registered, it receives from the cluster-head all sensed data in the cluster during the third phase. Possible new sensors are added to the cluster and the cluster-head is reelected during the fourth phase.

Yang et al. [10] propose a dynamic enclose cell (DEC) routing algorithm to decrease the routing overhead by constructing cells with sensor nodes in order to retain stable the WSN in high mobility. This protocol groups the nodes into cells and develops the routing path using the cells boundaries. When the nodes are moving, only the adjacent cells of the moving nodes are reconstructed. In this way, the negative impact of the node mobility is minimized. The DEC algorithm consists of four schemes: neighbor beacon exchange, cell discovery, cell routing path update and cell routing selection.

Arboleda et al. [11] propose a cluster-based routing (CBR) protocol for mobile WSNs using zone-base information and a cluster-like communication between nodes. It is based on two stages: route creation and route preservation. The first stage discovers a route between a source and a sink, but the second stage repairs the route when it is defective. The CBR protocol is based on the formation of non-overlapping square zones. Each node is placed in a zone and can obtain its zone ID based on its location parameters. The sensors in a common zone form a cluster and each cluster has one of the mobile nodes acting as cluster-head. This later acts as an aggregator node, receiving and forwarding messages to its neighbor cluster-heads, and maintains information about both the routes and the nodes in a zone.

Lambrou et al. [12] present a routing scheme in hybrid WSN that forwards packets to mobile nodes. The scheme objective is the delivery of event detection messages that contain information about position of the detected event in the sensor field. The routing of such messages can be easily achieved using a geographical routing based on greedy techniques towards a fixed base station. Moreover, this later easily requests information about a specific region or even a single static node using the position information.

Santhosh-Kumar et al. [13] propose an adaptive cluster-based routing (ACBR) scheme for mobile WSNs by including mobility as a new criterion for creation and maintenance of clusters. This work is considered as an improvement of the works proposed in [16, 17]. The ACBR protocol consists of two phases: set-up and steady-state.

Nasser et al. [14] propose a Zone-based Routing Protocol for mobile WSNs (ZoroMSN) based on zone construction, route maintenance and zone-head election. This protocol is efficient in WSN with low mobility of nodes, where clusters are formed using the mobility patterns of sensors. ZoroMSN acts as a hybrid routing protocol, where communication between nodes in a zone is proactive and between zone-heads towards the sink is reactive. The ZH is selected based on the mobility factor of each sensor in the zone, which is defined as the average number of times that a node moves from one zone to another during a given period of time.

Saad et al. [15] propose an energy efficient routing algorithm called Ellipse-Routing. Using a region-based routing, the proposed algorithm builds a virtual ellipse thanks to the source and destination position. So, only nodes within this ellipse can forward a message towards the destination. Then, the algorithm was extended in order to take into account errors in node location.

Although the above-resumed works play important roles in improving the performance of the geographical routing in mobile WSNs, the design of new routing solutions is still a challenging research area. Thus, the DGF mechanism is proposed in Section IV taking into account the mobility of nodes in WSNs. The DGF mechanism is associated with the well-known GPSR protocol and the obtained protocol is called GPSR-MS (GPSR with Mobile Sensors). The major difference between GPSR-MS and the above-summarized protocols includes the following aspects:

- The proposed GPSR-MS protocol operates without organizing the network into clusters, while the majority of existing protocols for mobile nodes are cluster-based where the maintenance consumes the limited resources of nodes.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

- In the existing cluster-based protocols, the greedy forwarding mode is not applied, while the GPSR-MS protocol is based on this scalable and efficient mode.
- Few of existing protocols are designed for WSNs with mobility of nodes. Therefore, the GPSR-MS protocol strengthens this class of protocols. Our objective is to maximize the packet delivery ratio with the minimum consumption of node resources.

III. NODE MOBILITY IMPACT ON GREEDY FORWARDING

In this section, we present the impact of nodes' mobility on the next forwarder selection. Majority of geographical routing protocols use greedy forwarding techniques to route packets in a WSN. To make their routing decision, they use only the locations of the forwarding candidate neighbors, the sender and the sink.

In greedy forwarding, the selected next-forwarder is the closest neighbor to the sink in term of distance based only on the nodes' location. But the mobility of nodes causes the problem of location information freshness inside the neighbors table of each sender node. This may result some routing decisions failures. This problem can be resolved by broadcasting location beacons. But when node mobility increases rapidly, the beaconing overhead grows also rapidly.

When the nodes move, the greedy forwarding mode does not often guarantee positive progression of packets towards their destination. Thus, when a sender node selects its next forwarder, this later may not be available because it moved. In the other hand, another node can come into the sender neighborhood, but it is not considered when selecting the next forwarder because it was not detected by the sender node. This situation has its importance when the non-detected node is the closest neighbor to the sink.

In addition, the moving direction and the moving speed of nodes may be the raison behind the obsolete table. Also, mobile nodes can repair holes that appear in a WSN due to their moving propriety. Then the greedy forwarding will use the shortest paths.

These greedy mode weaknesses induce packet losses, delivery delays and excessive energy consumption. Indeed, the use of only distance to select the intermediate forwarders has limits in dynamic environments caused by nodes' mobility. However, the use of periodic and frequent location beacons cannot resolve the problem because it creates packet collisions, overloads the network and consumes more energy. Consequently, some packets will be lost and other packets will be delayed. Therefore, the next-forwarder selection in a node must consider multiple metrics of its neighbors, such as the moving speed, the moving direction and the distance to the sink. The objective is to obtain a geographical routing protocol that maximizes the packet delivery ratio, minimizes the average path length and reduces the control packet overhead.

IV. PROPOSED DGF MECHANISM

To handle node mobility in mobile WSNs, the proposed DGF mechanism uses a new decision metric when selecting the next forwarder of the current packet. This metric considers the moving direction, the moving speed, and the distance to the sink of forwarding candidate neighbors of the sender node. The DGF mechanism supposes that each node moves with an angular variation according to the sink. However, each node i has an angle θ which is formed by neighbor n, sink s, and the horizontal axis passing by s, as shown in Figure 1. The moving direction of node n, between two recent times t_0 and t_1, is calculated by combining both its two last distances and angles to sink s. Neighbor n may become near to (or far from) the sink in terms of both distance and angle.

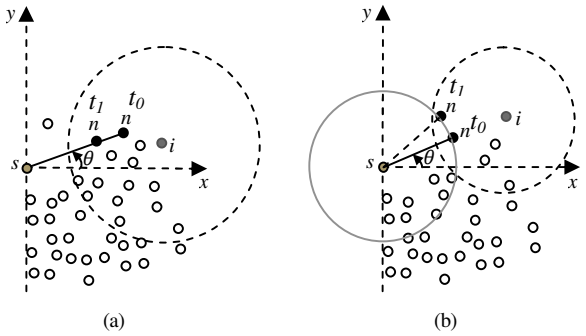


Figure 1. Neighbor moving: (a) s approaches sink s in term of distance and (b) n moves away from sink s in term of angle.

To show the neighbor direction evolution, the DGF mechanism combines the angle and distance parameters of neighbors. In Figure 1, the moving direction of neighbor n is calculated by node i using the two later parameters. The DGF mechanism operates in two main phases: neighbors' information update and next forwarder selection. Note that we suppose a WSN with a static sink and mobile nodes, each node knows its neighbors' positions, and sink's position thanks to the network initialization phase. Also, each node has a table (TABLE I) which contains information about its neighbors, such as location, moving speed and moving direction.

1) Neighbors' information update: Each node broadcasts periodically a 1-hop location beacon informing its neighbors about its geographical position. The period of this beacon can be fixed according to the nodes' moving speed. Thanks to these beacons, each node updates a local table containing information about all neighbors. We added to this table three new fields to record moving speed, angle and moving direction of each neighbor. TABLE I shows the structure of the neighbors table of a node. We also added a specific field into the location beacon, where the structure is given in TABLE II, to convey the moving speed of a node to all its neighbors.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

TABLE I. STRUCTURE OF A NEIGHBORS TABLE

Field	Mission/Content
ID	Identifier of a neighbor node
Position	Coordinates of a neighbor i (x_i, y_i)
Direction	Neighbor moving direction
Speed	Neighbor moving speed
Angle	Neighbor angle (θ) calculated according to the sink
ExpTime	Expire time of a neighbor in the table

When a node i receives a location beacon B from its neighbor n, it checks the existence of n in its neighbors table T. If node n does not exist, node i inserts information concerning n in T (TABLE I), else it calculates the old and new direction of neighbor n by using the formulas (1) and (2) respectively, where $DirT(n,s)$ represents the old direction calculated using T, $AT(n,s)$ is the old angle calculated using T, $DirB(n,s)$ is the new direction calculated using B and $AB(n,s)$ is the new angle calculated using B. The distances $DT(n,s)$ and $DB(n,s)$, between neighbor n and sink s, based on locations that are extracted from T, respectively from B, are given by the respective formulas (3) and (4). Note that $x_{n,T}$ and $y_{n,T}$ are locations of n in T, $x_{n,B}$ and $y_{n,B}$ are locations of n in B, x_s and y_s are locations of s in sender node i.

$$DirT(n,s) = DT(n,s) * AT(n,s) \quad (1)$$

$$DirB(n,s) = DB(n,s) * AB(n,s) \quad (2)$$

$$DT(n,s) = \sqrt{(x_{n,T} - x_s)^2 + (y_{n,T} - y_s)^2} \quad (3)$$

$$DB(n,s) = \sqrt{(x_{n,B} - x_s)^2 + (y_{n,B} - y_s)^2} \quad (4)$$

The angles $AT(n,s)$ and $AB(n,s)$, represented by θ in Figure 1, are calculated according to the trigonometric quadrant in which neighbor n is located by using the respective formulas (5) and (6) based on the $Arctg2(y,x)$ function of the C++ language. Once the above calculations are done by a node i, it updates all information concerning each neighbor n in its table T.

$$AT(n,s) = \quad (5)$$
$$\begin{cases} \frac{\pi}{2} - Arctg2(y_{n,T} - y_s, x_{n,T} - x_s); & \text{IF } x_{n,T} \geq x_s \text{ AND } y_{n,T} \geq y_s \\ \frac{\pi}{2} + Arctg2(y_{n,T} - y_s, x_{n,T} - x_s); & \text{IF } x_{n,T} < x_s \text{ AND } y_{n,T} \geq y_s \\ -\frac{\pi}{2} - Arctg2(y_{n,T} - y_s, x_{n,T} - x_s); & \text{IF } x_{n,T} < x_s \text{ AND } y_{n,T} < y_s \\ -\frac{\pi}{2} + Arctg2(y_{n,T} - y_s, x_{n,T} - x_s); & \text{IF } x_{n,T} \geq x_s \text{ AND } y_{n,T} < y_s \end{cases}$$

$$AB(n,s) = \quad (6)$$
$$\begin{cases} \frac{\pi}{2} - Arctg2(y_{n,B} - y_s, x_{n,B} - x_s); & \text{IF } x_{n,B} \geq x_s \text{ AND } y_{n,B} \geq y_s \\ \frac{\pi}{2} + Arctg2(y_{n,B} - y_s, x_{n,B} - x_s); & \text{IF } x_{n,B} < x_s \text{ AND } y_{n,B} \geq y_s \\ -\frac{\pi}{2} - Arctg2(y_{n,B} - y_s, x_{n,B} - x_s); & \text{IF } x_{n,B} < x_s \text{ AND } y_{n,B} < y_s \\ -\frac{\pi}{2} + Arctg2(y_{n,B} - y_s, x_{n,B} - x_s); & \text{IF } x_{n,B} \geq x_s \text{ AND } y_{n,B} < y_s \end{cases}$$

2) Next-forwarder selection: This phase aims to enhance the greedy mode of GPSR by handling parameters of the mobile nodes. Thus, we propose a new routing factor combining three parameters: 1) the distance $DT(n,s)$ between neighbor n and sink s, 2) the moving direction $ABDir(n,s)$ of the neighbor n and 3) the moving speed $Speed(n)$ of neighbor n. When a node i has to send a packet to sink s, by using a greedy forwarding, it selects from its neighbors table node n having the smallest $DBFactor(n,s)$ given by Formula (7), where the direction $ABDir(n,s)$ is given by Formula (8). Note that this direction is calculated using the formulas (1) and (2). When $ABDir(n,s)$ is equal to 1 then n is static. When it is greater than 1 then n approaches the sink. When it is less than 1 then n moves away from the sink.

$$DBFactor(n,s) = \frac{DT(n,s) * ABDir(n,s)}{Speed(n)} \quad (7)$$

$$ABDir(n,s) = \frac{DirT(n,s)}{DirB(n,s)} \quad (8)$$

TABLE II. STRUCTURE OF A LOCATION BEACON

Field	Mission/Content
ID	Identifier of the node that sent a beacon
Position	Location of the node that sent a beacon
Speed	Moving speed of the node that sent a beacon

V. PERFORMANCE EVALUATION

We first implemented and evaluated the traditional GPSR protocol using ns2 [18] with mobility of nodes. Then we associated the proposed DGF mechanism with GPSR and evaluated in same conditions the resulting protocol (GPSR-MS). Since GPSR can handle mobility of nodes by reducing the location beacon period, we evaluate performance of this protocol under four values of this period (2ms, 3ms, 4ms and 5ms) and obtained the results which are shown in the graphs as GPSR(2), GPSR(3), GPSR(4) and GPSR(5), respectively. This period is set to 5ms for the GPSR-MS protocol.

In our simulations, we used a terrain 600m×600m with 350 mobile sensors deployed randomly. Then they move according to Random Waypoint Model (RWM) with a random speed in [5-20] m/s to simulate the mobility in realistic environments. The sink is placed at the center of the terrain and 12 sources are selected randomly. Each source generates one CBR flow with a rate increased gradually from 1 to 12 p/s. For each rate and at the end of the simulation time, we measure the packet delivery ratio, the control packet overhead, the average path length and the network energy consumption per delivered packet. Table III gives the parameters settings used in our simulations.

Compared to GPSR in Figure 2, our GPSR-MS protocol achieves a better packet delivery ratio, especially when the rate is less than 5p/s. The number of packets dropped in

GPSR is important when a beaconing period is large (5ms). Figure 3 shows a good performance of GPSR-MS in term of average path length compared to GPSR. This is due to our DGF mechanism which dynamically selects as forwarders the neighbors that move toward the sink.

TABLE III. SIMULATION ENVIRONMENT SETTINGS

Bandwith	200 Kbps
Playload	32 Bytes
Terrain	600m × 600m
Number of nodes	350 nodes
Node Speed	Random in [5-20] m/s
Node Radio Range	40 m
MAC Layer	802.11
Radio Layer	RADIO-NONNOISE
Propagation Model	TWO-RAY
Simulation Time	224 sec
Mobility Model	RWM Version 1

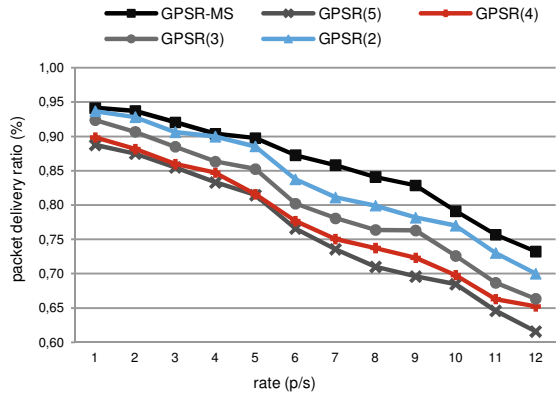


Figure 2. Performance in delivering data packets.

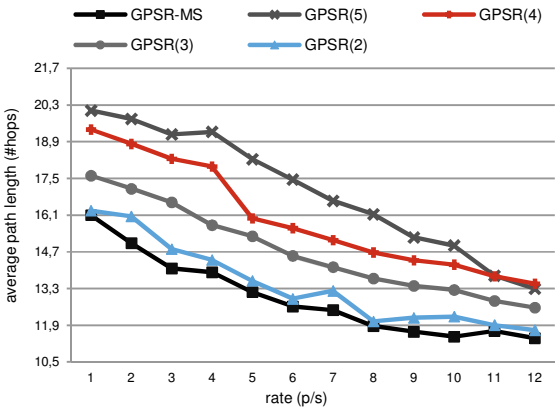


Figure 3. Performance in reducing the paths length.

Note that when the location beacon is not large (2ms), the average path length is reduced in GPSR because tables of neighbors are frequently updated. Consequently, GPSR generates more location beacons which overload the sensor network (Figure 4) and consume excessive energy of nodes (Figure 5). On the other hand, GPSR-MS delivers more data packets, generates less control packets and manages correctly the limited energy of nodes.

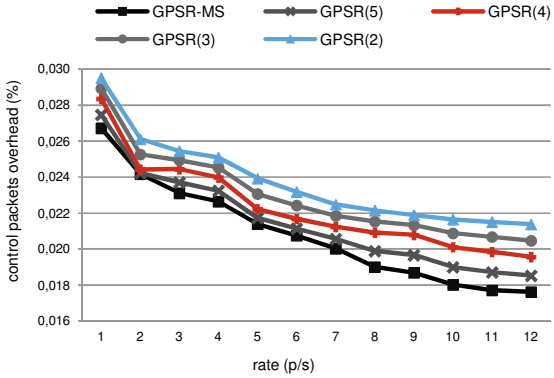


Figure 4. Performance in reducing control packets.

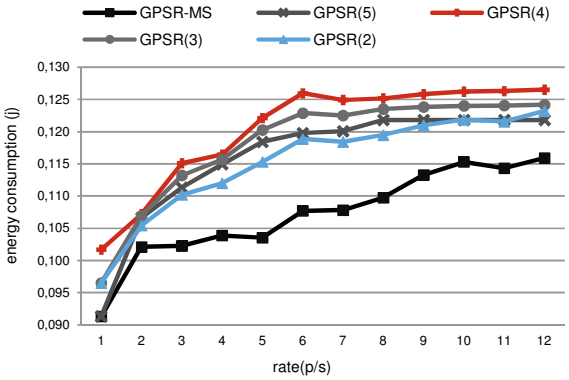


Figure 5. Performance in economizing energy of nodes.

VI. CONCLUSION

Existing geographical schemes using greedy forwarding in mobile WSNs still have problems as mentioned above (Section III). To contribute on solving these problems, we have proposed the DGF mechanism that handles mobility of nodes in WSNs. It is simple to implement, saves the network resources and could be associated with various geographical routing protocols. The merit of our proposal is that the current packet is forwarded to the best neighbor node in terms of distance, moving direction and moving speed according to the static sink. We have associated the DGF mechanism with the well-known GPSR protocol and the resulting protocol, called GPSR-MS, has achieved good performance compared to different versions of the original GPSR. Indeed, GPSR-MS delivers more packets, broadcasts less control packets, uses the shortest routing paths and economizes much energy of nodes. Our future work will evaluate performance of GPSR-MS with the group mobility concept.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

REFERENCES

[1] S.A. Munir, B. Ren, W. Jiao, B. Wang, D. Xie, and J. Ma, "Mobile wireless sensor network: Architecture and enabling technologies for ubiquitous computing", Proc. of the 21st Int'l Conference on Advanced Information Networking and Applications Workshops, pp. 113–120, Ontario, Canada, May 21-23, 2007.

[2] B. Liu, P. Brass, O. Dousse, P. Nain, and D. Towsley, "Mobility improves coverage of sensor networks", Proc. of the 6th ACM international symposium on Mobile ad hoc networking and computing, pp. 300-308, Illinois, USA, May 25-27, 2005.

[3] B. Karp and H. Kung, "GPSR: Greedy perimeter stateless routing for wireless networks," Proc. of the ACM/IEEE Conference on Mobile Computing and Networking, pp. 243-254, Boston, Massachusetts, USA, August 6-11, 2000.

[4] Y. Xu, J. Heidemann, and D. Estrin, "Geography-informed Energy Conservation for Adhoc Routing", Proc. of the 7th Annual ACM/IEEE International Conference on Mobile Computing and Networking, pp. 70-84, Rome, Italy, July 16-21, 2001.

[5] T. He, J.A. Stankovic, C. Lu, and T. Abdelzaher, "A Spatiotemporal Communication Protocol for Wireless Sensor Networks," IEEE Transactions on Parallel and Distributed Systems, Vol. 16(10), pp. 995-1006, October 2005.

[6] Q. Fang, J. Gao, and L.J. Guibas, "Locating and bypassing holes in sensor networks," IEEE Mobile Networks and Applications, Vol. 11(2), pp. 187–200, April 2006.

[7] F. Ye, H. Luo, J. Cheng, S. Lu, and L. Zhang, "A Two-Tier Data Dissemination Model for Large-scale Wireless Sensor Networks", Proc. of the Mobile Computing and Networks Conference, pp.148-159, Georgia, USA, Sept. 23-28, 2002.

[8] K. Fodor and A. Vidacs, "Efficient Routing to Mobile Sinks in Wireless Sensor Networks", Proc. of the 2nd International Workshop on Performance Control in WSNs (PWSN), pp. 1–7, Austin, Texas, USA, October 23, 2007.

[9] Y.H. Wang, K.F. Huang, P.F. Fu, and J.X. Wang, "Mobile Sink Routing Protocol with Registering in Cluster-Based WSNs", Proc. of the 5th International Conference on Ubiquitous Intelligence and Computing, pp. 527-535, Oslo, Norway, June 23-25, 2008.

[10] Y. Yang, L. Dong-Hyun, P.K. Myong-Soon, and I.H. Peter, "Dynamic Enclose Cell Routing in Mobile Sensor Networks", Proc. of the Asia-Pacific Software Engineering Conference (APSEC), pp.736-737, Busan, Korea, Nov. 30 – Dec. 3, 2004.

[11] M. Liliana, C. Arboleda, and N. Nidal, "Cluster-based Routing Protocol for Mobile Sensor Networks", Proc. of the 3rd International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks, pp. 24-28, Canada, August 7–9, 2006.

[12] C.G. Panayiotou, T. Theofanis, and P. Lambrou, "A Survey on Routing Techniques supporting Mobility in Sensor Networks", Proc. of the 5th International Conference on Mobile Ad-hoc and Sensor Networks, pp. 78–85, China, Dec. 14-16, 2009.

[13] G.S. Kumar, A. Sitara, and K.P. Jacob "An adaptive cluster based routing scheme for mobile wireless sensor networks", Proc. of the 2nd International Conference on Computing, Communications and Networking Technologies, pp. 1-5, Karur, India, July 29-31, 2010.

[14] N. Nasser, A. Al-Yatama, and K. Saleh, "Mobility and Routing in Wireless Sensor Networks," Proc. of the 24th Annual IEEE Canadian Conference on Electrical and Computer Engineering, Niagara Falls, pp. 573-578, Canada, May 8-11, 2011.

[15] C. Saad, A. Benslimane, J. Champ, and J.C. Konig, "Ellipse routing: A Geographic Routing Protocol for Mobile Sensor Networks with uncertain positions", Proc. of the 2nd Int'l Conference on Future Generation Communication and Networking, pp. 1–5, Sanya, China, December 13-15, 2008.

[16] D.S. Kim and Y.J. Chung, "Self-Organization Routing Protocol Supporting Mobile Nodes for Wireless Sensor Network", Proc. of the First Int'l Multi-Symposiums on Computer and Computational Sciences, pp. 622-626, Hangzhou, China, June 20-24, 2006.

[17] G.S. Kumar, M.V. Paul and K.P. Jacob, "Mobility Metric based LEACH-Mobile Protocol", Proc. of 16th International Conference on Advanced Computing and Communications, pp. 248–53, Chennai, India, December 14-17, 2008.

[18] Collaboration between researchers at UC Berkeley, LBL, USC/ISI, and Xerox PARC, "The ns Manual", on line at: http://www.isi.edu/nsnam/ns/, last consultation in December March 2012.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Bit Error Rate for Complex SSC/MRC Combiner in the Presence of Nakagami-*m* Fading

Dragana Krstić, Mihajlo Stefanović,
Department of Telecommunications
Faculty of Electronic Engineering, University of Niš
Niš, Serbia
dragana.krstic@elfak.ni.ac.rs

Petar Nikolić
Tigartyres,
Pirot, Serbia
nikpetar@gmail.com

***Abstract*—The complex Switch and Stay Combining/Maximal Ratio Combining (SSC/MRC) combiner is considered in this paper. The system output signal at two time instants is observed in the presence of Nakagami-*m* fading at the input. Both of combiners, SSC and MRC, are with two branches. The probability density function (PDF) at the output of the complex combiner is obtained and the bit error rate (BER) for the case of binary phase swift keying (BPSK) modulation is determined. The obtained results are shown graphically. It was pointed out the improvement of using complex SSC/MRC combiner relative to classical MRC and SSC combiners at one time instant.**

Keywords-Bit error rate, Probability density function; Complex SSC/MRC combiner; Nakagami-m fading; two time instants

I. INTRODUCTION

The fading, the occurrence of variation of instantaneous value of the received signal envelope, is one of the very important factors of signal quality derogation at the reception. Many urban and vehicular communication systems are subjected to fading caused by multipath propagation due to reflection, refraction and scattering by buildings and other large structures [1]. The received signal is, thus, a sum of different signals that arrive via different propagation paths.

Several statistical models have been used in the literature to describe the fading envelope of the received signal [2]-[8]. The Rayleigh and Rician distributions are used to characterize the envelope of faded signals over small geographical areas or sort term fades, while the log-normal distribution is used when much wider geographical areas are involved. A more versatile statistical model, however, is Nakagami-*m*-distribution [7], which can model a variety of fading environments including those modelled by the Rayleigh and one-sided Gaussian distributions. Also the log-normal and Rician distributions may be closely approximated by the Nakagami distribution in some ranges of mean signal values [2]. The fit between Nakagami and Rician distributions may not be very good when the signal-noise ratio (SNR) is large, but is very accurate for low SNR values. Furthermore, the Nakagami distribution is more flexible and more accurately fit experimental data for many physical propagation channels then the log-normal and Rician distributions [2], [3]. Although the Nakagami model fits experimental data around the mean or median, it does not fit very well in the tails of the distribution. For this reason, some researchers question the use of the Nakagami model [4]; however, there is continued interest in modelling a variety of propagation channels with the Nakagami distribution [2], [3]. Moreover, in [9] Braun and Dersch give a detailed derivation of the Nakagami distribution and show that it is quite appropriate to model multipath fading in the mobile radio channel.

There are several ways to reduce the impact of fading on system performances. The goal is to achieve this without increasing the signal power and channel capacity. The diversity reception techniques are used extensively in fading radio channels to reduce the effects of fading on system performances [6], [10], and [11], including both fixed terminals and mobile communication systems.

In order to gain significantly from the use of diversity, there must be a sufficient degree of statistical independence in the fading of the received signal in each of the diversity branches. The assumption of statistical independence between the diversity channels is valid only if they are sufficiently separated [2]. In mobile radio systems the signals at the mobile station become decorrelated as the antenna separation (or frequency separation) increases, giving rise to diversity. If the antennas are crowded then such diversity conditions may be violated [8]. In space diversity systems an antenna separation of 30 to 50 wavelengths is typically required to obtain correlation coefficients strictly between zero and one-third, in which case, for a two-channel maximal-ratio system in a Rayleigh-fading environment, the effect of correlation may be ignored [10].

However, there are other cases of practical interest where the assumption of statistical independence is not valid. When the Nakagami channel is studied, the analysis is usually limited to the dual-branch diversity system [12]. A long time ago Al-Hussani and Al-Bassion studied the effect of correlation on the performance of a dual-branch maximal-ratio combiner for the correlated Nakagami-fading channel. They found that for the Nakagami-fading environment and for a worst case fading condition and identical signal-to-noise ratio (SNR) in each of the two branches, the performance difference between a single channel and the two channels system increases from 3 to 24 dB as the correlation coefficient decreases from unity to zero.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

II. RELATED WORK

In diversity systems multiple copies of the same signal are sending. They are combined in different ways in order to obtain as larger as possible signal to noise ratio. There are some kinds of diversity combining schemes.

Maximal-Ratio Combining (MRC) is the optimal combining scheme [13]. In this combiner signals from all inputs are summed. Because MRC requires cognition of the channel fading parameters, it is the most complicated and expensive combining model [14], [15].

Equal Gain Combining (EGC) is next and then Selection Combining (SC) and Switch and Stay Combining (SSC), with lower performances. These combining models are simpler and cheaper and they are very often implemented in practice whereas SC and SSC combining models do not require signal cophasing and fading envelope evaluation [16].

SSC is simplification of the system complexity, but with losing in quality. In this model, the receiver selects an antenna until its value drops below predetermined threshold. Then, the receiver switches to other antenna and remains for the next time slot, no matter the channel quality of that antenna is above or below the threshold. In the literature, mainly dual SSC schemes have been analyzed [17], [18].

By the authors' knowledge in the new open literature, except from papers published by this group of authors, there are no papers that treat these problems by sampling in the two time instants. We derived the expression for the joint probability density function of the SSC combiner output signal in the presence of different fading distributions (Nakagami, log-normal, Hoyt) in two time instants [19]-[21]. Based on these joint PDFs we made the performance analysis of SSC/SC combiner at two time instants in the presence of Rayleigh and log-normal fading in [22], [23]. The bit error rates for SSC/MRC combiner at two time instants in the presence of log-normal, Rayleigh and Hoyt fading we determined in [24]-[26], respectively.

In this paper the probability density function and the bit error rate of the SSC/MRC combiner output signal in the presence of Nakagami-*m* fading, with sampling signals at two time instants for one time slot, will be observed. The system is more complex then classical MRC and SSC systems at one time instant, but with better performances. That means that bit error rate can be increased and transmit power can be reduced comparing to classical systems.

This paper is organized as follows: II Section gives related works; III Section describes the complex SSC/MRC system model and the process of obtaining the probability density functions and the bit error rate of the SSC/MRC combiner output signal at two time instants. Sections IV presents numerical results obtained for performances introduced in previous section. Finally, the main results of the paper are presented in V Section as conclusions.

III. COMPLEX SSC/MRC COMBINER MODEL

The model of the SSC/MRC combiner with two inputs considering in this paper is shown in Fig. 1.

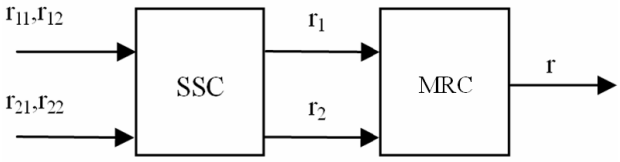


Figure 1. Complex dual SSC/MRC combiner.

We consider the SSC/MRC combiner with two branches at two time instants. The signals at the inputs at SSC combiner are r_{11} and r_{21} at first time moment and they are r_{12} and r_{22} at the second time moment. The output signals at SSC part are r_1 and r_2. The indexes for the input signals are: first index is the number of the branch and the other signs time instant observed. For the output signals, the index represents the time instant observed. After determining the output signals at SSC combiner r_1 and r_2, they become the inputs at MRC combiner and the overall output signal is r.

The joint probability density function of correlated signals r_1 and r_2 at the output of SSC combiner at two time inputs with Nakagami-*m* distribution and for the same parameters, is obtained in [19].

For $r_1 < r_T, r_2 < r_T$ it is:

$$p^1{}_{r_1 r_2}(r_1, r_2) =$$

$$= P_1 \sum_{k=0}^{\infty} \frac{2(r_2)^{2m_1+2k-1}}{k!\Gamma(m_1)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_1}{\Omega_1}\right)^{m_1+k} e^{-\frac{m_1 r_2^2}{\Omega_1(1-\rho)}} \gamma\left(\frac{m_1}{\Omega_1(1-\rho)} r_T{}^2, m_1+k\right) \cdot$$

$$\cdot \sum_{k=0}^{\infty} \frac{2(r_1)^{2m_2+2k-1}}{k!\Gamma(m_2)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_2}{\Omega_2}\right)^{m_2+k} e^{-\frac{m_2 r_1^2}{\Omega_2(1-\rho)}} \gamma\left(\frac{m_2}{\Omega_2(1-\rho)} r_T{}^2, m_2+k\right) +$$

$$+ P_2 \sum_{k=0}^{\infty} \frac{2(r_2)^{2m_2+2k-1}}{k!\Gamma(m_2)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_2}{\Omega_2}\right)^{m_2+k} e^{-\frac{m_2 r_2^2}{\Omega_2(1-\rho)}} \gamma\left(\frac{m_2}{\Omega_2(1-\rho)} r_T{}^2, m_2+k\right) \cdot$$

$$\cdot \sum_{k=0}^{\infty} \frac{2(r_1)^{2m_1+2k-1}}{k!\Gamma(m_1)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_1}{\Omega_1}\right)^{m_1+k} e^{-\frac{m_1 r_1^2}{\Omega_1(1-\rho)}} \gamma\left(\frac{m_1}{\Omega_1(1-\rho)} r_T{}^2, m_1+k\right) \tag{1}$$

For $r_1 \geq r_T, r_2 < r_T$:

$$p^2{}_{r_1 r_2}(r_1, r_2) = P_1 \frac{2m_2{}^{m_2} r_2{}^{2m_2-1}}{\Omega_2{}^{m_2}\Gamma(m_2)} e^{-\frac{m_2 r_2{}^2}{\Omega_2}} \cdot$$

$$\cdot \sum_{k=0}^{\infty} \frac{2(r_1)^{2m_1+2k-1}}{k!\Gamma(m_1)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_1}{\Omega_1}\right)^{m_1+k} e^{-\frac{m_1 r_1^2}{\Omega_1(1-\rho)}} \gamma\left(\frac{m_1}{\Omega_1(1-\rho)} r_T{}^2, m_1+k\right) +$$

$$+ P_1 \sum_{k=0}^{\infty} \frac{2(r_2)^{2m_1+2k-1}}{k!\Gamma(m_1)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_1}{\Omega_1}\right)^{m_1+k} e^{-\frac{m_1 r_2^2}{\Omega_1(1-\rho)}} \gamma\left(\frac{m_1}{\Omega_1(1-\rho)} r_T{}^2, m_1+k\right) \cdot$$

$$\cdot \sum_{k=0}^{\infty} \frac{2(r_1)^{2m_2+2k-1}}{k!\Gamma(m_2)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_2}{\Omega_2}\right)^{m_2+k} e^{-\frac{m_2 r_1^2}{\Omega_2(1-\rho)}} \gamma\left(\frac{m_2}{\Omega_2(1-\rho)} r_T{}^2, m_2+k\right) +$$

$$+ P_2 \frac{2{m_1}^{m_1} {r_2}^{2m_1-1}}{{\Omega_1}^{m_1}\Gamma(m_1)} e^{-\frac{m_1 {r_2}^2}{\Omega_1}} \cdot$$

$$\cdot \sum_{k=0}^{\infty} \frac{2(r_1)^{2m_2+2k-1}}{k!\Gamma(m_2)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_2}{\Omega_2}\right)^{m_2+k} e^{-\frac{m_2 r_1^2}{\Omega_2(1-\rho)}} \gamma\left(\frac{m_2}{\Omega_2(1-\rho)} {r_T}^2, m_2+k\right) +$$

$$+ P_2 \sum_{k=0}^{\infty} \frac{2(r_2)^{2m_2+2k-1}}{k!\Gamma(m_2)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_2}{\Omega_2}\right)^{m_2+k} e^{-\frac{m_2 r_2^2}{\Omega_2(1-\rho)}} \gamma\left(\frac{m_2}{\Omega_2(1-\rho)} {r_T}^2, m_2+k\right) \cdot$$

$$\cdot \sum_{k=0}^{\infty} \frac{2(r_1)^{2m_1+2k-1}}{k!\Gamma(m_1)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_1}{\Omega_1}\right)^{m_1+k} e^{-\frac{m_1 r_1^2}{\Omega_1(1-\rho)}} \gamma\left(\frac{m_1}{\Omega_1(1-\rho)} {r_T}^2, m_1+k\right) \tag{2}$$

For $r_1 < r_T, r_2 \geq r_T$:

$$p^3{}_{r_1 r_2}(r_1, r_2) = P_1 \gamma\left(\frac{m_1}{\Omega_1} {r_t}^2, m_1\right) \cdot$$

$$\cdot \frac{4(r_1 r_2)^{m_2}}{\Gamma(m_2)(1-\rho)\rho^{(m_2-1)/2}} \left(\frac{m_2}{\Omega_2}\right)^{m_2+1} I_{m-1}\left(\frac{2m_2\sqrt{\rho}\, r_1 r_2}{\Omega_2(1-\rho)}\right) e^{-\frac{m_2\left(r_1^2+r_2^2\right)}{\Omega_2(1-\rho)}} +$$

$$+ P_1 \sum_{k=0}^{\infty} \frac{2(r_2)^{2m_1+2k-1}}{k!\Gamma(m_1)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_1}{\Omega_1}\right)^{m_1+k} e^{-\frac{m_1 r_2^2}{\Omega_1(1-\rho)}} \gamma\left(\frac{m_1}{\Omega_1(1-\rho)} {r_T}^2, m_1+k\right) \cdot$$

$$\cdot \sum_{k=0}^{\infty} \frac{2(r_1)^{2m_2+2k-1}}{k!\Gamma(m_2)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_2}{\Omega_2}\right)^{m_2+k} e^{-\frac{m_2 r_1^2}{\Omega_2(1-\rho)}} \gamma\left(\frac{m_2}{\Omega_2(1-\rho)} {r_T}^2, m_2+k\right) +$$

$$+ P_2 \gamma\left(\frac{m_2}{\Omega_2} {r_t}^2, m_2\right) \cdot \frac{4(r_1 r_2)^{m_1}}{\Gamma(m_1)(1-\rho)\rho^{(m_1-1)/2}} \left(\frac{m_1}{\Omega_1}\right)^{m_1+1} \cdot$$

$$\cdot I_{m-1}\left(\frac{2m_1\sqrt{\rho}\, r_1 r_2}{\Omega_1(1-\rho)}\right) e^{-\frac{m_1\left(r_1^2+r_2^2\right)}{\Omega_1(1-\rho)}} +$$

$$+ P_2 \sum_{k=0}^{\infty} \frac{2(r_2)^{2m_2+2k-1}}{k!\Gamma(m_2)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_2}{\Omega_2}\right)^{m_2+k} e^{-\frac{m_2 r_2^2}{\Omega_2(1-\rho)}} \gamma\left(\frac{m_2}{\Omega_2(1-\rho)} {r_T}^2, m_2+k\right) \cdot$$

$$\cdot \sum_{k=0}^{\infty} \frac{2(r_1)^{2m_1+2k-1}}{k!\Gamma(m_1)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_1}{\Omega_1}\right)^{m_1+k} e^{-\frac{m_1 r_1^2}{\Omega_1(1-\rho)}} \gamma\left(\frac{m_1}{\Omega_1(1-\rho)} {r_T}^2, m_1+k\right) \tag{3}$$

For $r_1 \geq r_T, r_2 \geq r_T$:

$$p^4{}_{r_1 r_2}(r_1, r_2) = P_1 \frac{4(r_1 r_2)^{m_1}}{\Gamma(m_1)(1-\rho)\rho^{(m_1-1)/2}} \left(\frac{m_1}{\Omega_1}\right)^{m_1+1} \cdot$$

$$\cdot I_{m-1}\left(\frac{2m_1\sqrt{\rho}\, r_1 r_2}{\Omega_1(1-\rho)}\right) e^{-\frac{m_1\left(r_1^2+r_2^2\right)}{\Omega_1(1-\rho)}} + P_1 \frac{2{m_2}^{m_2} {r_2}^{2m_2-1}}{{\Omega_2}^{m_2}\Gamma(m_2)} e^{-\frac{m_2 r_2^2}{\Omega_2}} \cdot$$

$$\cdot \sum_{k=0}^{\infty} \frac{2(r_1)^{2m_1+2k-1}}{k!\Gamma(m_1)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_1}{\Omega_1}\right)^{m_1+k} e^{-\frac{m_1 r_1^2}{\Omega_1(1-\rho)}} \gamma\left(\frac{m_1}{\Omega_1(1-\rho)} {r_T}^2, m_1+k\right) +$$

$$+ P_1 \gamma\left(\frac{m_1}{\Omega_1} {r_t}^2, m_1\right) \cdot \frac{4(r_1 r_2)^{m_2}}{\Gamma(m_2)(1-\rho)\rho^{(m_2-1)/2}} \left(\frac{m_2}{\Omega_2}\right)^{m_2+1} \cdot$$

$$\cdot I_{m-1}\left(\frac{2m_2\sqrt{\rho}\, r_1 r_2}{\Omega_2(1-\rho)}\right) e^{-\frac{m_2\left(r_1^2+r_2^2\right)}{\Omega_2(1-\rho)}} +$$

$$+ P_1 \sum_{k=0}^{\infty} \frac{2(r_2)^{2m_1+2k-1}}{k!\Gamma(m_1)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_1}{\Omega_1}\right)^{m_1+k} e^{-\frac{m_1 r_2^2}{\Omega_1(1-\rho)}} \gamma\left(\frac{m_1}{\Omega_1(1-\rho)} {r_T}^2, m_1+k\right) \cdot$$

$$\cdot \sum_{k=0}^{\infty} \frac{2(r_1)^{2m_2+2k-1}}{k!\Gamma(m_2)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_2}{\Omega_2}\right)^{m_2+k} e^{-\frac{m_2 r_1^2}{\Omega_2(1-\rho)}} \gamma\left(\frac{m_2}{\Omega_2(1-\rho)} {r_T}^2, m_2+k\right) +$$

$$+ P_2 \frac{4(r_1 r_2)^{m_2}}{\Gamma(m_2)(1-\rho)\rho^{(m_2-1)/2}} \left(\frac{m_2}{\Omega_2}\right)^{m_2+1} I_{m-1}\left(\frac{2m_2\sqrt{\rho}\, r_1 r_2}{\Omega_2(1-\rho)}\right) e^{-\frac{m_2\left(r_1^2+r_2^2\right)}{\Omega_2(1-\rho)}} +$$

$$+ P_2 \frac{2{m_1}^{m_1} {r_2}^{2m_1-1}}{{\Omega_2}^{m_1}\Gamma(m_1)} e^{-\frac{m_1 r_2^2}{\Omega_2}} \cdot$$

$$\cdot \sum_{k=0}^{\infty} \frac{2(r_1)^{2m_2+2k-1}}{k!\Gamma(m_2)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_2}{\Omega_1}\right)^{m_2+k} e^{-\frac{m_2 r_1^2}{\Omega_1(1-\rho)}} \cdot \gamma\left(\frac{m_2}{\Omega_1(1-\rho)} {r_T}^2, m_2+k\right) +$$

$$+ P_2 \gamma\left(\frac{m_2}{\Omega_2} {r_t}^2, m_2\right) \cdot \frac{4(r_1 r_2)^{m_1}}{\Gamma(m_1)(1-\rho)\rho^{(m_1-1)/2}} \left(\frac{m_1}{\Omega_1}\right)^{m_1+1} \cdot$$

$$\cdot I_{m-1}\left(\frac{2m_1\sqrt{\rho}\, r_1 r_2}{\Omega_1(1-\rho)}\right) e^{-\frac{m_1\left(r_1^2+r_2^2\right)}{\Omega_1(1-\rho)}} +$$

$$+ P_2 \sum_{k=0}^{\infty} \frac{2(r_2)^{2m_2+2k-1}}{k!\Gamma(m_2)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_2}{\Omega_2}\right)^{m_2+k} e^{-\frac{m_2 r_2^2}{\Omega_2(1-\rho)}} \gamma\left(\frac{m_2}{\Omega_2(1-\rho)} {r_T}^2, m_2+k\right) \cdot$$

$$\cdot \sum_{k=0}^{\infty} \frac{2(r_1)^{2m_1+2k-1}}{k!\Gamma(m_1)} \left(\frac{\rho}{1-\rho}\right)^k \left(\frac{m_1}{\Omega_1}\right)^{m_1+k} e^{-\frac{m_1 r_1^2}{\Omega_1(1-\rho)}} \gamma\left(\frac{m_1}{\Omega_1(1-\rho)} {r_T}^2, m_1+k\right) \tag{4}$$

where m_i and Ω_i are parameters of Nakagami-*m* distribution, ρ is correlation coefficient and r_t is the threshold of the decision for SSC combiner.

Total conditional signal value at the MRC combiner output, for equally transmitted symbols of *L* branch MRC receiver, is given by

$$r = \sum_{l=1}^{L} r_l \tag{5}$$

For coherent binary signals the conditional BER $P_b\left(e \middle| \{r_l\}_{l=1}^{L}\right)$ is given by [13]:

$$P_b(e|\{r_l\}_{l=1}^{L}) = Q\left(\sqrt{2gr}\right) \quad (6)$$

where Q is the one-dimensional Gaussian Q-function [1]

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-t^2/2} dt \quad (7)$$

Gaussian Q-function is defined as [27]

$$Q(x) = \frac{1}{\pi} \int_0^{\pi/2} \exp\left(-\frac{x^2}{2\sin^2\phi}\right) d\phi \quad (8)$$

If alternative representation of Gaussian-Q function is used, the conditional BER can be expressed as

$$P_b(e|\{r_l\}_{l=1}^{L}) = \frac{1}{\pi} \int_0^{\pi/2} \exp\left(-\frac{gr}{\sin^2\phi}\right) d\phi = \frac{1}{\pi} \int_0^{\pi/2} \prod_{l=1}^{L} \left(-\frac{gr_l}{\sin^2\phi}\right) d\phi \quad (9)$$

The unconditional BER can be obtained by averaging the multichannel conditional BER over the joint PDF of the signals at the input of MRC combiner

$$P_b(e) = \underbrace{\int_0^{\infty}\int_0^{\infty}\cdots\int_0^{\infty}}_{L} P_b\left(\{r_l\}_{l=1}^{L}\right) \prod_{l=1}^{L} p_{r_1,r_2,\ldots r_L}(r_1, r_2, \ldots r_L) dr_1 dr_2 \ldots dr_L \quad (10)$$

Substituting (9) in (10), $P_b(e)$ is obtained as

$$P_b(e) = \underbrace{\int_0^{\infty}\int_0^{\infty}\cdots\int_0^{\infty} \frac{1}{\pi} \int_0^{\pi/2} \prod_{l=1}^{L} \left(-\frac{gr_l}{\sin^2\phi}\right) d\phi}_{L} p_{r_1,r_2,\ldots r_L}(r_1, r_2, \ldots r_L) dr_1 dr_2 \ldots dr_L \quad (11)$$

For dual branch MRC combiner, $P_b(e)$ is

$$P_b(e) = \underbrace{\int_0^{\infty}\int_0^{\infty}\cdots\int_0^{\infty} \frac{1}{\pi} \int_0^{\pi/2} \prod_{l=1}^{L} \left(-\frac{gr_l}{\sin^2\phi}\right) d\phi}_{L} p_{r_1,r_2,\ldots r_L}(r_1, r_2, \ldots r_L) dr_1 dr_2 \ldots dr_L \quad (12)$$

Substituting (1-4) in (12), $P_b(e)$ of SSC/MRC combiner can be obtained as:

$$P_b(e) = \frac{1}{\pi} \int_0^{r_t}\int_0^{r_t}\int_0^{\pi/2} dr_1 dr_2 d\phi \left(-\frac{gr_1}{\sin^2\phi}\right)\left(-\frac{gr_2}{\sin^2\phi}\right) p^1{}_{r_1 r_2}(r_1, r_2) +$$

$$+ \frac{1}{\pi} \int_{r_t}^{\infty}\int_0^{r_t}\int_0^{\pi/2} dr_1 dr_2 d\phi \left(-\frac{gr_1}{\sin^2\phi}\right)\left(-\frac{gr_2}{\sin^2\phi}\right) p^2{}_{r_1 r_2}(r_1, r_2) +$$

$$+ \frac{1}{\pi} \int_0^{r_t}\int_{r_t}^{\infty}\int_0^{\pi/2} dr_1 dr_2 d\phi \left(-\frac{gr_1}{\sin^2\phi}\right)\left(-\frac{gr_2}{\sin^2\phi}\right) p^3{}_{r_1 r_2}(r_1, r_2) +$$

$$+ \frac{1}{\pi} \int_{r_t}^{\infty}\int_{r_t}^{\infty}\int_0^{\pi/2} dr_1 dr_2 d\phi \left(-\frac{gr_1}{\sin^2\phi}\right)\left(-\frac{gr_2}{\sin^2\phi}\right) p^4{}_{r_1 r_2}(r_1, r_2) \quad (13)$$

IV. NUMERICAL RESULTS

The bit error rate curves, for different types of combiners and correlation parameters, are presented in Fig.2 and 3. It is assumed that both inputs have the same channel parameters. r_t is the optimal threshold for the SSC decision [8]:

$$r_t = \frac{\Gamma(m+1/2)}{\Gamma(m)} \left(\frac{\Omega}{m}\right)^{1/2} \quad (14)$$

The BER family curves for one channel receiver, for MRC combiner at one time instant and for SSC/MRC combiner at two time instants for uncorrelated case, and also for very strong correlation, are shown in Fig. 2 versus different distribution parameter.

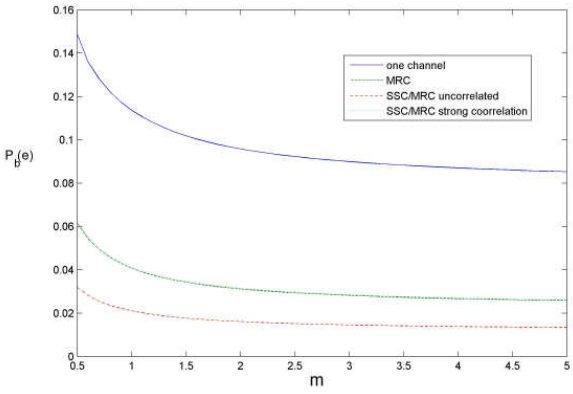


Figure 2. Bit error rate for different types of combiners versus parameter m , for $\Omega=0.5$

One can see from this figure that SSC/MRC combiner has significant better performances for uncorrelated case then MRC combiner at one time instant. For $\rho = 1$ the BER of SSC/MRC combiner follows the results for MRC combiner.

It is obvious that using of complex SSC/MRC combiner results in better performance of the system because the BER for uncorrelated SSC/MRC combiner decrease for about 50% regarding MRC combiner.

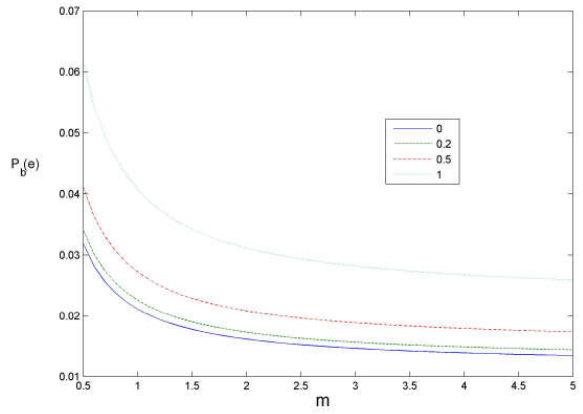


Figure 3. Bit error rate for SSC/MRC combiner versus parameter m for $\Omega=0.5$, for different values of ρ

The influence of correlation to the outage probability of complex SSC/MRC combiner is presented in Fig. 3. The benefits of using this type of combiner increases with decreasing of correlation between input signals. It is apparent that there is no economic justification for the use of complex SSC/MRC combiner in the case of strong correlations between input signals.

V. CONCLUSIONS

The SSC and MRC are simple and frequently used techniques for signal combining in diversity systems. The probability density function of complex dual SSC/MRC combiner output signal, at two time instants, is determined in this paper. The bit error probability is expressed based on it.

The system performances deciding by two samples can be determine by the joint probability density function of the SSC combiner output signal at two time instants and putting them as inputs of MRC combiner. The obtained results are shown graphically. The performance improvement of SSC/MRC combiner at two time instants, comparing with classical SSC and MRC combiners, is described.

ACKNOWLEDGMENT

This work has been funded by the Serbian Ministry for Science under the projects TR-32052, III-44006 and TR-33035.

REFERENCES

[1] V. A. Aalo, “Performance of maximal-ratio diversity systems in a correlated Nakagami fading environment," IEEE. Trans. Commun., vol. 43, pp. 2360-2369, Aug. 1995.

[2] H. Suzuki, “A statistical model for urban radio propagation,” IEEE Trans. Commun., vol. COM-25, pp. 673-680, July 1977.

[3] U. Charash, “Reception through Nakagami fading multipath channels with random delays,” IEEE Trans. Commun., vol. COM-27, pp. 657-670, Apr. 1979.

[4] S. Stein, “Fading channel issues in system engineering,” IEEEJ. Selecr. Areas Commun., vol. SAC-5, pp. 6849, Feb. 1987.

[5] C. Loo, “Statistical model for a land mobile satellite link,” IEEE Trans. Veh. Technol. vol. VT-34, pp. 122-127, Aug. 1985.

[6] W. C. Jakes, Mobile Communication Engineering, New York: Wiley, 1974.

[7] M. Nakagami, “The iff-distribution-A general formula of intensity distribution of rapid fading,” in Starisrical Methods in Radio Wave Propagation, W. G. Hoffman, Ed. Oxford, England : Pergamon, 1960.

[8] G. L. Turiner al., “A statistical model of urban multipath propagation,” IEEE Trans. Veh. Techno/. vol. VT-21, pp. 1-8, Feb. 1972.

[9] W. R. Braun and U. Dersch, “A physical mobile radio channel mode,” IEEE Trans. Veh. Technol., vol. 40, pp. 472-482, Feb. 1991.

[10] D. G. Brennan, “Linear diversity combining techniques,” in Proc. IRE, vol. 47, pp. 1075-1102, June 1959.

[11] M. Schwartz, W. R. Bennett, and S. Stein, Communicarion Systems and Techniques, New York: McGraw-Hill, 1966.

[12] E. AI-Hussaini and A. AI-Bassiouni, “Performance of MRC diversity systems for the detection of signals with Nakagami fitding,” IEEE Trans. Commun., vol. COM-33, pp. 1315-1319, Dec. 1985.

[13] M. K. Simon, M. S. Alouni, Digital Communication over Fading Channels, Second Edition, Wiley-Interscience, A John Wiley&Sons, Inc., Publications, New Jersey, 2005.

[14] M. Stefanović, N. Kapicinović, M. Bandjur, "Moments of the MRC and EGC combiner output", Electronic and electrical engineering, no.1 (73), pp. 59-63, 2007.

[15] G. Popovic, S. Panic, J. Anastasov, M. Stefanovic, P. Spalevic, “Cooperative MRC diversity over Hoyt fading channels”, Electrical Review, vol. 87 no. 12, pp. 150-152, Dec. 2011.

[16] Đ. V. Banđur, M. V. Banđur, M. Stefanović, ”Output Signal Characteristics of a Post-detection EGC Combiner with Two ASK Input Signals in the Presence of Fading and Gaussian Noise”, Electronics and Electrical Engineering (Elektronika ir elektrotechnika), no. 2(82), T180 Telecommunications engineering, pp. 41-44, 2008.

[17] N. Sekulović, M. Stefanović, D. Drača, A. Panajotović, M. Zdravković, “Switch and stay combining diversity receiver in microcellular mobile radio system”, Electrical Review (Przeglad Elektrotechniczny), vol. 86, no. 2, pp. 346-350, Dec. 2010.

[18] A.Panajotović, M. Stefanović, D. Drača, “Cochannel Interference Effect on BEP Performance of SSC Receiver in Correlated Rician Fading”, Journal of the Franklin Institute, vol. 347, no. 7, pp. 1242-1252, 2010.

[19] D. Krstić, P. Nikolić, M. Matović, A. Matović, M. Stefanović, “The Joint Probability Density Function of the SSC Combiner Output Signal in the Presence of Nakagami-m fading”, The Fourth International Conference on Wireless and Mobile Communications, ICWMC 2008 and ICCGI 2008, [including the workshop ComP2P 2008], Athens/Vouliagmeni, Greece, July 27-August 1 2008, pp. 409-416, ISBN 978-0-7695-3274-5

[20] D. Krstić, P. Nikolić, F. Destović, M. Stefanović, „The Joint Probability Density Function of the SSC Combiner Output Signal in the Presence of Log-Normal Fading”, Electronics and Electrical Engineering, (Elektronika ir Elektrotechnika), ISSN 1392-1215, No. 3(109), pp. 11-16, 2011.

[21] D. Krstić, P. Nikolić, G. Stamenović, M. Stefanović, “The joint probability density function of the SSC Combiner Output Signal at two Time Instants in the Presence of Hoyt Fading”, The Seventh International Conference on Wireless and Mobile Communications, ICWMC 2011, 19-24. June, Luxembourg

[22] P. Nikolić, D. Krstić, M. Milić, and M. Stefanović, “Performance Analysis of SSC/SC Combiner at Two Time Instants in The Presence of Rayleigh Fading”, Frequenz. Volume 65, Issue 11-12, Pages 319–325, ISSN (Online) 2191-6349, ISSN (Print) 0016-1136, November/2011, http://www.reference-global.com/doi/abs/10.1515/FREQ.2011.048

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

[23] M. Stefanović, P. Nikolić, D. Krstić, V. Doljak, "Outage probability of the SSC/SC combiner at two time instants in the presence of lognormal fading", Przeglad Elektrotechniczny (Electrical Review), ISSN 0033-2097, R. 88 NR 3a/2012, pp.237-240, march 2012

[24] D. Krstić, P. Nikolić, G. Stamenović, M. Stefanović "Bit error rate for SSC/MRC Combiner at Two Time Instants in The Presence of log-normal Fading", Facta Universitatis. Series Automatic Control and Robotics, ISSN 1820-6417, Vol.10, No 1, pp. 83 – 95, UDC 621.396.94 621.395.38 519.724, 2011.

[25] D. Krstić, P. Nikolić, S. Panić, V. Doljak, "The Bit Error Rate for Complex SSC/MRC Combiner in the Presence of Rayleigh Fading", The third International Conference on Information and Communication Systems (ICICS 2012), Irbid, Jordan, April 3- 5, 2012.

[26] D. Krstić, P. Nikolić, G. Stamenović, M. Stefanović, "The Bit Error Rate for Complex SSC/MRC Combiner at Two Time Instants in the Presence of Hoyt Fading", Article 8321 is accepted for publication as Tele12v5n12 in International Journal on Advances in Telecommunications, 2012.

[27] J.W. Craig, "A new, simple and exact result for calculating the probability of error for two-dimensional signal constellations," IEEE MILCOM'91 Conf. Rec., Boston, MA, pp. 25.5.1–25.5.5.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Performance Evaluation of a WiMAX Network Using Smart Antennas Through System in the Loop OPNET Simulations

Şerban Georgică Obreja,
University POLITEHNICA of Bucharest
Bucharest, Romania
e-mail: serban@radio.pub.ro

Alexey Baraev
Create-Net
Trento, Italy
e-mail: alexey.baraev@create-net.org

Irinel Olariu,
University POLITEHNICA of Bucharest
Bucharest, Romania
e-mail: irinel.olariu@elcom.pub.ro

Eugen Borcoci
University POLITEHNICA of Bucharest
Bucharest, Romania
e-mail: eugen.borcoci@elcom.pub.ro

***Abstract*—Worldwide Interoperability for Microwave Access (WiMAX) is one of the newest technologies developed for broadband wireless networks, which offers high data rate and high flexibility for the radio resource management. Adding smart antenna support and developing new scheduling algorithms will make WiMAX an attractive solution for the next generation wireless networks. In this paper, a basic simulation testbed for performance evaluation of a WiMAX network using smart antenna is proposed. It is based on OPNET simulation tool and uses System-in-the-Loop function to interconnect the simulated system with a real network, for a better functional and performance evaluation. The simulation results showed that the higher antenna gain on the beam direction and the narrow beam will result in a reduced level of interference and a higher throughput in the WiMAX network.**

Keywords - WiMAX networks; smart antenna; OPNET; System-in-the-Loop.

I. INTRODUCTION

The IEEE 802.16 technology and WiMAX-based systems constitute an attractive solution for metropolitan and rural areas [1]. It offers high capacity links, and based on the relay technology introduced in 802.16 j standard it can also provide high coverage too. Scheduling the radio resources in a relay based topology is a very challenging task. To cope with interference issues in such topologies, directional smart antennas can be introduced to obtain increase in performance, while keeping the transmitting power at the same level as for omnidirectional antennas case.

In this paper, the performance of a WiMAX network, which uses smart antennas at the Base Station and Subscriber Station, is evaluated. A basic testbed was built for this purpose. It is based on OPNET simulator and uses the System-in-the-Loop function to interconnect the simulated system with a real network, for a better functional and performance evaluation [2][3]. The testbed was developed in the framework of the SMART-Net FP7 project, which aimed to investigate the use of smart antennas in Wireless Mesh Networks mainly based on WiMAX and WiFi technologies. The project proposed efficient scheduling algorithms, to enhance the capacity and to provide scalability, reliability and robustness for such a system [2][4]. Inside this project, performances evaluation based on both simulation and real life experimental platform has been conducted. In order to increase the accuracy of the evaluation platform, the cooperation between the simulated network and the real life platform has been achieved by coupling them using the System- in-the-loop OPNET function. This paper presents the experimental results for evaluating the smart antennas integration on standard OPNET WiMAX nodes [5].

This paper is organized as follows. The second section is a short description of the SMART-Net system and of the smart antennas. The third section is dedicated to the testbed structure and the simulation scenarios. The fourth section provides the simulation results. The last section presents conclusions and guidelines for future work.

II. SMART-NET SYSTEM

A. *Smart-Net features*

Smart-Net solution for Broadband Wireless Access is based on multimode devices with smart antennas support. These devices are interconnected in a partial mesh topology, which has a central point, Smart Gateway (SMG), acting as a gateway linked to the backhaul networks. The other nodes of the network are either SMART Stations (SMS) or SMART Relays (SMR). A SMR is an operator's equipment, which is specifically used to forward data traffic to the users, allowing coverage extension and cooperative diversity, while a SMS is a subscriber station, that also enables data transfer for other users based on the service provider policy [1] [4].

Using omnidirectional antennas in wireless networks create inherently interference, which decrease the capacity of the system. A significant capacity increase could be obtained by using smart antennas. They feature a directivity that can be controlled by higher levels protocols (Layer 2, Layer 3) in the network node, allowing its orientation towards the destination node, and thus reducing interference.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

The Smart-Net project introduced smart-antenna support on WiMAX equipments and developed some algorithms for scheduling and routing in a multihop relay based WiMAX mesh network [4][6][7][8]. To validate the proposed solutions two testbeds were developed during the project [5]. First is a real life testbed consisting of WiMAX equipments with smart antennas. The WiMAX equipments used in the testbed are produced by Thales Company and the smart antennas are produced by Plasma Antennas Company. Both are members of the Smart-Net project. The second is a simulation platform, which was developed using the OPNET network simulator. The smart antennas were modeled in the OPNET and integrated with the simulated WiMAX nodes. Also, it was proposed by the project to combine the real life and simulated testbeds to obtain more significant results of the proposed system's performances. For the testbeds interconnection the System-in-the-Loop (SITL) function provided by OPNET was used. Such approach is presented in [9], where SITL is used to evaluate WiFi wireless networks performances. In this paper a similar approach is used to evaluate the smart antennas integration on the WiMAX nodes and the smart antennas performances.

B. *Smart antennas*

Smart antennas are systems, which intelligently combine multiple antenna elements with signal-processing capability to optimize its radiation and reception patterns automatically [4]. They have a certain number of fixed high gain beams with low sidelobes, which minimize interference both on transmit and receive, without using complex adaptive nulling algorithms. Low sidelobe multi-beam antennas have the advantage over adaptive systems in that they suppress a very large number of interferers in a consistent, predictable way. Adaptive antennas systems are limited by their degrees of freedom (e.g., number of radios), their adaptation time and might not work well when the signal of interest is at the edge of the receiver's sensitivity. However, they potentially have the advantage of allowing the suppression of interfering signals that are close in angle (within a beamwidth) to the source-of-interest. When receiving, the adaptive smart antennas can maximize the sensitivity in the direction of the desired signal and minimize the sensitivity towards interfering sources.

For reasons of cost and consistency of performance, common smart antennas are switched or selectable multi-beam antennas, requiring only a single radio. These antennas have multiple fixed beams, and the system switches very rapidly between these beams.

For the SMART-Net project, two types of switched multi-beam antennas, capable of WiMAX operation, have been designed and implemented [2] [4].

- An active, 12 beam cylindrical array antenna with omni-mode
- A passive 9 beam planar array antenna with sectoral mode.

The active 12-beam cylindrical antenna with 360° coverage has been selected to be most suitable for mesh and nomadic Point to Multipoint operation. It has typical ranges of up to 20 km, depending on the modulation rate. The passive 9 beam planar antenna, with its narrow beams, has been selected to be most suitable for medium range backhaul and relay operations. A representation of both antennas, suitable for inclusion within OPNET, has also been provided, but as simulated data.

Besides the multibeam antennas, a switching algorithm is used to choose the appropriate beam among the available antenna beams. This algorithm is based on a learning interval in which, based on SINR, the best beam is chosen for each destination. Based on the decision took by the selection algorithm, when a smart node (a node equipped with smart antennas) needs to communicate with another smart node, the beam with the best SINR is used. Because the best beam is decided in the learning phase, the switch operation is very fast, a few nanoseconds. Some performance degradation is expected in the mobile nodes case, because the learning interval lasts a few milliseconds. In this paper, only evaluation of smart antenna on fixed WiMAX nodes is presented. For mobile nodes, the smart antenna integration is not ready.

III. SIMULATION INFRASTRUCTURE AND SCENARIOS FOR SMART ANTENNAS

A. *Simulated testbed infrastructure*

The testbed infrastructure consists of a simulated WiMAX network, which is interconnected with real devices in order to introduce real time traffic in the simulation (Figure 1). The System-in-the-Loop is an OPNET facility, which allows real time communication between real and simulated parts of the network [2][8][10]. By using SITL, OPNET simulation exchanges the packets between simulations and real networks in real-time. The SITL gateway represents an external device through which the simulation exchanges the packets; the WinPcap library is used to route those packets selected by user defined filter, from an Ethernet network adaptor, to the simulation process. The real time requirements are introducing hard constraints on the simulation platform's hardware.

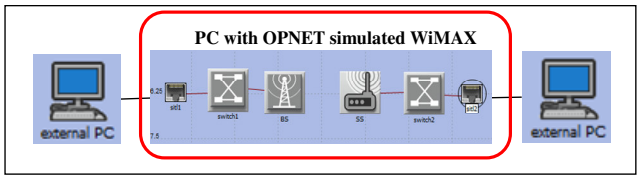


Figure 1. Simulated testbed infrastructure.

The simulation runs in real-time and exchanges packets with the external hardware via an Ethernet link. The requirement of using Ethernet link between the real devices and the SITL gateway introduces limitation in developing joint real and simulated wireless network scenarios. Joint scenarios for evaluating scheduling and routing algorithms for wireless mesh networks are not possible with SITL.

B. *Scenarios for smart antenna performance evaluation*

The basic topology of the simulated WiMAX network used for evaluation of smart antennas performances is

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

presented in Figure 1. A real-to-real SITL topology is used for these scenarios. It consists of a single WiMAX link, between a Base Station and a fixed Subscriber Station, which is concatenated with Ethernet links at both ends. These Ethernet links are used to interconnect, via SITL gateways, the WiMAX simulated network with the external stations, which are both acting as real time streaming server and player. The smart antennas were installed on the simulated WiMAX nodes. Introducing smart antenna support on standard WiMAX nodes requires modifying the radio transceiver pipeline stages. The beam selection algorithm was introduced in the pipeline stages together with the 12 beam cylindrical array and 9 beam planar array antenna models.

The following two scenarios were created in order to evaluate the smart antennas integration in standard WiMAX nodes using the OPNET SITL tool. The first one uses the topology given in Figure 1, which consists of a single WiMAX link. Both BS and SS are equipped with standard WiMAX nodes (with omnidirectional antennas) initially. The capacity of a standard WiMAX link is determined. In a second phase, the omnidirectional antennas are replaced with smart antennas. The scenario is run again to determine the capacity of the WiMAX link for smart antenna case. The second scenario aimed to determine the interference level in a WiMAX network when the smart antennas are used taken as reference the interference generated by omnidirectional antennas. For this purpose, near the WiMAX link, used to carry the real time traffic, a small WiMAX network (one Base Station- BS, with several Subscriber Stations- SS) is placed. The nodes of this network are used to generate interference on the main WiMAX link. The scenario is run with both omni and smart antennas installed on the nodes used to generate interference.

IV. Simulation Results For Smart Antenna Performance Evaluation

A. *Simulation components and parameters*

In this section, the smart antennas performance evaluation results obtained using the simulation testbed will be presented. The simulation platform consists from the following components as shown in Figure 1: two PCs used one as real traffic generator and the second as player; a third PC, with OPNET installed, is used to simulate the WiMAX network. The PC with OPNET must be a performant one in order to run the simulations in real time. The hardware simulation platform used for the experiments is based on Intel XEON-quad core processor, running at 2.66GHz, with 6GB RAM. Windows 7 and OPNET version 16 software are installed on it. Two 1GB Ethernet cards are used to interconnect the simulation PC with the external PCs. In the first scenario a direct WiMAX link, between a Base Station and a Subscriber Station, is simulated in OPNET. Both WiMAX nodes are standard nodes equipped with omnidirectional or smart antennas. The WiMAX physical parameters are: 20MHz bandwidth, 2048 subcarriers, 10.94 kHz subcarrier frequency spacing, symbol duration of 102.86µs, frame duration of 5ms. The antenna gain is set at 15 dB and the receiver sensitivity was set at -100dB.

A first test suite aimed to evaluate if there are any limitations introduced by the SITL interface in the WiMAX performances. For this purpose, the free space propagation model was chosen, multipath channel model was disabled, and the distance between BS and SS was set at around 200m. With these parameters for the radio channel, the transmission is done in very good conditions. The Iperf application was used to generate UDP traffic at a rate of 50 Mbps. Several modulations and coding rates were configured for WiMAX physical layer: QAM64 3/4, 2/3, 1/2; QAM 16 3/4 and 1/2. The data throughput obtained on the output SITL interface for different modulations and coding rates are presented in Figure 2. The data throughput through the WiMAX link, measured at the output of the simulated network and at the destination PC, reaches the maximum link capacity for the given WiMAX parameters. The obtained results show that the SITL interfaces and the real-time constraints imposed to the simulation does not affect the network performances.

As can be seen from Figure 2, the throughputs obtained through the WiMAX link in each case are closed to the values indicated by the standard.

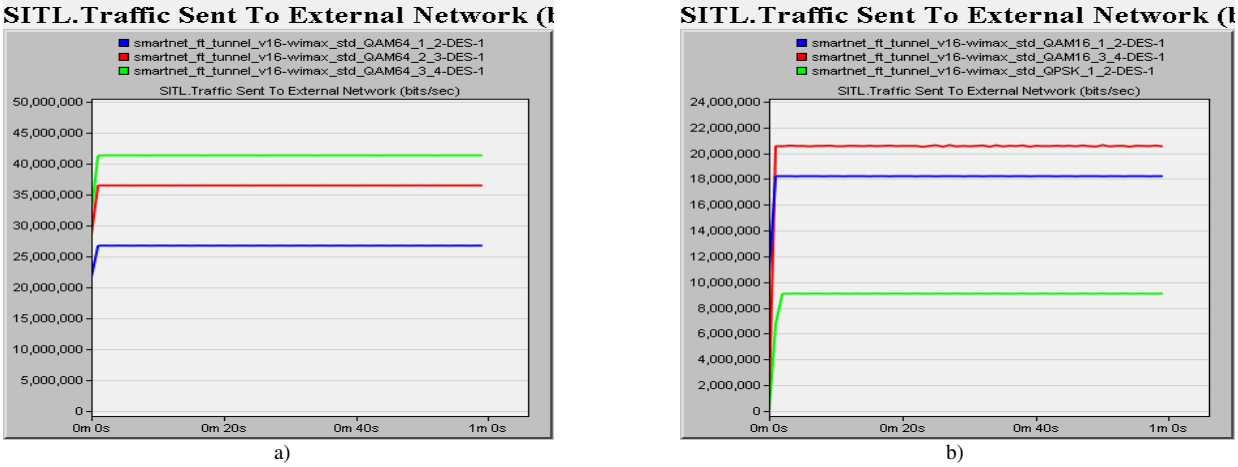


Figure 2. Data throughput obtained on the output SITL interface for:a) QAM 64 modulation 1/2, 2/3, 3/4 b) QAM16 1/2, 3/4 and QPSK 1/2

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

B. *Basic evaluation of link capacity for WiMAX nodes equipped smart antennas*

In this scenario, a movie with a rate around 2.5 Mbps is streamed from the streaming server through the simulated WiMAX link. On the same link and in the same direction (downlink) it is transmitted a noise UDP traffic with the rate of 5Mbps. All the flows are transmitted as Best Effort. A total of around 7.5 Mbps throughput is transmitted on the downlink. The WiMAX physical parameters are: 20MHz bandwidth, 2048 subcarriers, 10.94 kHz subcarrier frequency spacing, symbol duration of 102.86 µs, frame duration of 5ms. The antenna gain is set at 15 dB. QPSK 1/2 modulation and coding scheme were selected, and the planar 9 multibeam smart antennas were installed on WiMAX nodes.

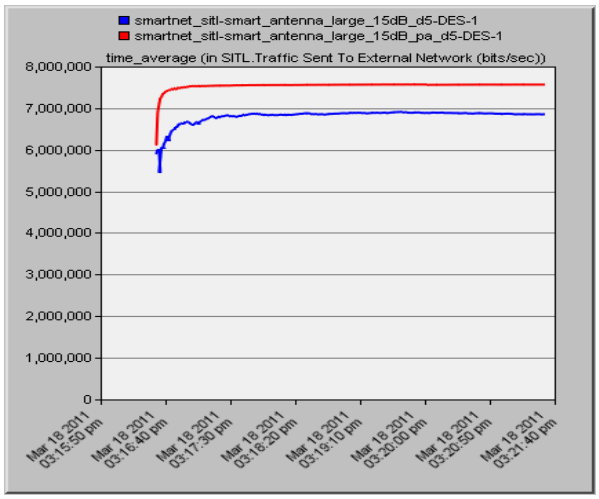


Figure 3. Wimax throughput – omni beam and smart-antenna; blue – omnidirectional antenna; red – smart antenna

Both the omni and smart antennas scenarios were repeated by varying the distance between the BS and SS. As it was expected the capacity of the WiMAX link decreases while the distance between nodes is increased. The Wimax capacity decrease is illustrated also by the perceptual evaluation of the movie quality and by the throughput statistics measured on the WiMAX link. In all experiments performed, the WiMAX link capacity was similar or better than in case of omnidirectional antenna as is illustrated in the Figures 3. It presents the throughput obtained in the omni-beam and smart antenna case when the BS and SS are at the same distance and in the same positions. The higher throughput curve corresponds to the smart antenna scenario. For small distances between the BS and SS node the throughput is the same. When the distance is increased the difference between the throughputs obtained in each case is increased – the higher throughput being obtained when using smart antennas. A perceptual evaluation was performed using the real time movie, which was sent through the simulated networks together with the UDP noise traffic. By subjectively observing the movie quality at the output of the simulator, it was compared the WiMAX network behavior in the omnidirectional and smart antennas cases. In all the scenarios, the movie quality was the same or better for the case when the smart antennas were used on WiMAX nodes compared with the case of omnidirectional antennas usage on WiMAX nodes.

C. *Interference evaluation for omni and smart antennas cases*

A second scenario was built to evaluate the interference level in a WiMAX network with nodes equipped with smart antennas, and to compare it with the interference generated by WiMAX nodes equipped with omnidirectional antennas. In order to evaluate the interference level, near the WiMAX link from the previous scenarios was placed a small WiMAX network with one BS and several subscribers. This second network is used to generate interference on the link, which is used to send the real time traffic flows.

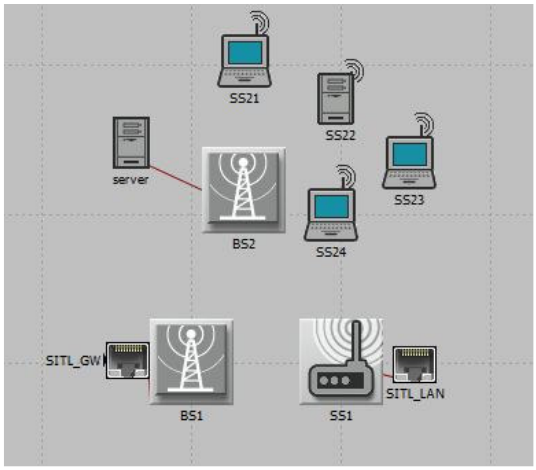


Figure 4. Interference evaluation scenario topology

The scenario is run both with the interference network enabled and disabled. The behavior of the main WiMAX link is evaluated for both cases: with and without interference. The same experiment is repeated for WiMAX nodes equipped with smart antennas. With this scenario, the interference generated by a WiMAX network with omni or smart antenna equipped nodes can be evaluated at a node level from the SNR statistics. The expected result is that the interference amount will be smaller when smart antennas are used by the WiMAX nodes.

The scenario topology is shown in Figure 4. The interference network consists from the server node, BS2 node, and the SS21, SS22, SS23, SS24 nodes. Local video traffic is generated between the interference nodes. The real time traffic is generated with Iperf application at a rate of 20 Mbps. For the WiMAX channel the ITU Pedestrian A multipath channel model and Pedestrian A pathloss model were selected. The other WiMAX parameters are identical with the ones used previous scenarios.

First experiments are performed using nodes equipped with omnidirectional antennas. The same experiment is run initially with the nodes, which generate interference, disabled, and then with all the nodes enabled.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

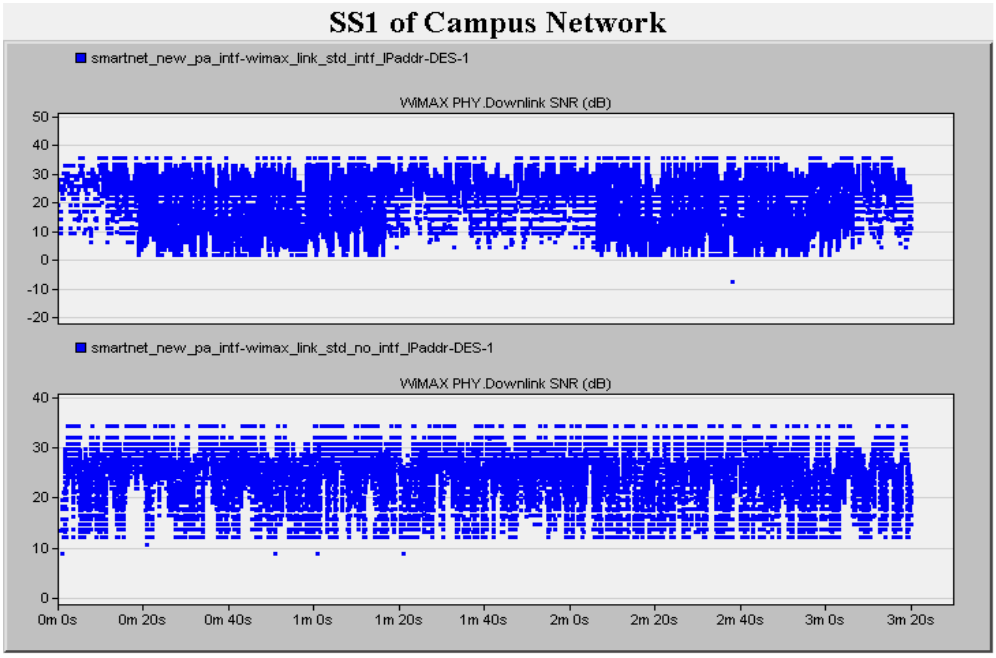


Figure 5. Downlink SINR values for the signal received by the SS1 node: **up-** interference nodes enabled; **down-** interference nodes disabled

The results are shown in the Figures 5, 6 and 7. In these figures, the downlink SNR measured at the SS1 station and the traffic sent in the interference network are presented. One can see that, when there is traffic in the interference network (Figure 6), the SNR level measured by the SS1 on the affected link is decreasing with almost 10 dB (Figure 5). This is caused by the radio signals coming from the interference WiMAX network. In Figure 7 the packets lost statistic measured at a node, SS21, in the second WiMAX network is shown. The packets lost is caused by the interference, which is generated by the WiMAX nodes BS1 and SS1 while transmitting the real time traffic. One can see that a lot of packets are lost because of the SNR degradation.

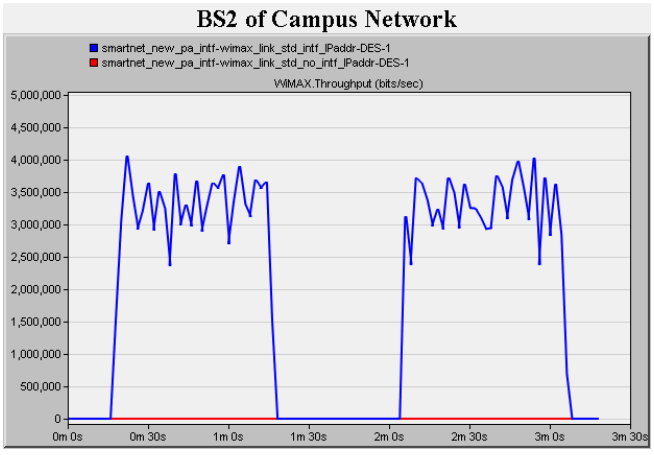


Figure 6. The throughput at the BS2 node –traffic generating interference at SS1 node

In both Figures 6 and 7, the blue curve corresponds to the results obtained when the nodes, which generate interference, are enabled, while the red curved illustrates the statistics obtained when the interference generating nodes are disabled.

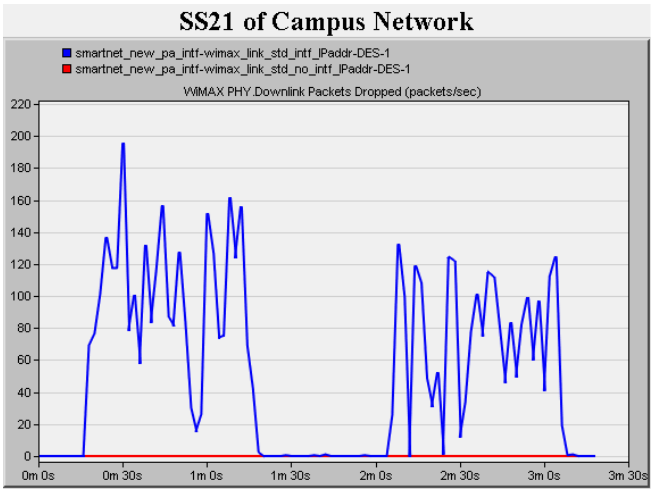


Figure 7. Packets lost at SS21 node – caused by the interference generated by the BS1-SS1 WiMAX link

A second test suite is done using nodes equipped with smart antennas. The topology and all other parameters were kept unchanged. A similar traffic pattern was used for this test suite. Local video traffic was generated between the interference nodes, and real time traffic was generated with *Iperf* application at a rate of 20 Mbps. The first experiment was played with the interference network activated. The

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

SNR measured at SS1 node, when the interference nodes are enabled, is shown in Figure 8. The SNR is similar with the SNR measured by SS1 when the nodes are equipped with omnidirectional antennas and the interference nodes are disabled. This shows that the level of the interference signals is low when smart antennas are used. Also, one can see that, the higher values of the SNR, in the scenario with nodes equipped with smart antennas, are with more than 10 dB greater than the SNR in the scenario with omnidirectional antennas. This is a consequence of the narrow beams used by the SS1 and BS1 antennas, which cause a smaller amount of interference due to multipath propagation.

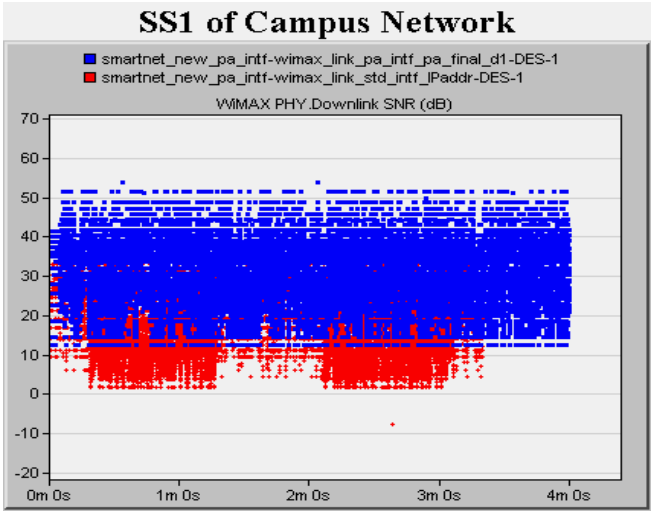


Figure 8. Downlink SNR at SS1 node - interference network enabled
rede – omnidirectional antenna; blue – smart antenna

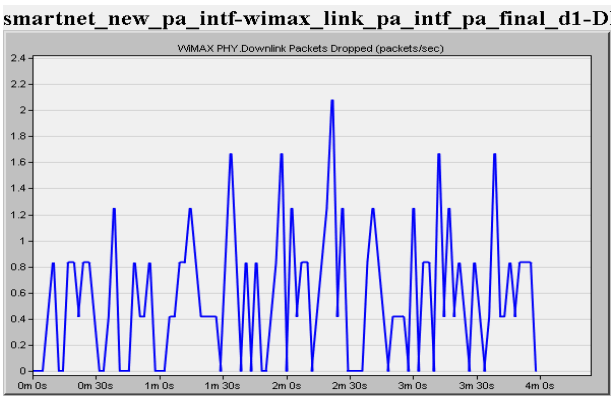


Figure 9. The number of downlink packets dropped at SS1 node

In Figure 9, the packet lost statistic for the node SS1 is presented when the interference network is enabled. There is a mean of less than one packet/s loss, which is much better than the loss obtained when the omnidirectional antennas are used. All these results demonstrate that using directional smart antennas on WiMAX nodes one can increase the capacity and coverage of the network. Because the beam switching is very fast, a few nanoseconds order, these smart antennas can be used at both Base Station and Subscriber Station for fixed and nomadic scenarios.

V. CONCLUSION AND FUTURE WORKS

This paper presents an evaluation, using the OPNET simulator, of a basic WiMAX network with nodes equipped with smart antennas. It presents basic scenarios and experiments for evaluation of smart antennas integration on WiMAX nodes and the obtained performances. The simulations results presented in this paper illustrate that, for fixed WiMAX nodes, the usage of the proposed smart antennas will bring a significant performance gain, expressed in terms of capacity and coverage. Because of the fast beam switching, these smart antennas can be used successfully in fixed scenarios. The smart antenna behavior on mobile WiMAX nodes is a subject of further research. The main issue for mobility is to develop a tracking algorithm capable of detecting in real time the best beam to be used to reach the mobile node.

ACKNOWLEDGMENT

This work was supported by the EU FP7 project SMART-Net, no 223937, by the Romanian UEFISCSU PN-2 RU-TE Project no. 18/12.08.2010 and by the EU and Romanian Govern EXCEL project - POSDRU/89/1.5/S/62557.

REFERENCES

[1] IEEE Standard 802.16-2004, Air Interface for Fixed Broadband Wireless Access Systems, wirelessman.org/pubs/ 80216-2004.html [retrieved: March, 2012].

[2] S. Wendt, E. Borcoci, and A. Tonnerre. Project Deliverable D2.1 a: Target scenario requirements and specifications, ICT FP7 SMART-Net project. https://www.ict-smartnet.eu [retrieved: March, 2012].

[3] http://www.opnet.com/training/network_rd/abstracts/mod_SITL-module.html [retrieved: March, 2012].

[4] S. Wendt, F. Kharrat-Kammoun, E. Borcoci, R. Cacoveanu, R. Lupu, and D. Hayes. Project Deliverable, “D2.4b: Network architecture and system specification,” ICT FP 7 SMART-Net project, October 2010, https://www.ict-smartnet.eu [retrieved: February, 2012].

[5] E. Borcoci, M. Constantinescu, S. Obreja, A.Baraev, T.Rasheed, and D. E. Meddour. Project Deliverable, “D4.4: System level simulation analyses and performance measures,” ICT FP7 SMART-Net project, May 2011.

[6] M. Mostafavi, E. Hamadani, et al., Project Deliverable, “D3.2b: Performance analysis of efficient routing protocols for multimode mesh networks,” ICT FP 7 SMART-Net project, December 2010,

[7] F. K. Kammoun, D. E. Meddour, A. Baraev, T. Rasheed, E. Borcoci, A. Enescu, and S. Ciochina. Project Deliverable, “D4.1: Large scale simulation testbed specifications,” ICT FP7 SMART-Net project. January 2009.

[8] R. Kortebi, D. E. Meddour, Y. Gourhant, and N. Agoulmine, “SINR-based routing in multi-hop wireless networks to improve VoIP applications support,” Consumer Communications and Networking Conference, CCNC 2007, Las Vegas,USA, pp491–496, 10.1109/CCNC.2007.103 7.

[9] J. Mohorko, M. Fras, and Ž. Čučej, ”Real time “system-in-the-loop” simulation of tactical networks,” 16th International Conference on Software, Telecommunicati ons and Computer Networks, SoftCOM 2008, pp. 105-108, 25-27 Sept. 2008

[10] S. G. Obreja, I. Olariu, A. Baraev, and E. Borcoci, “Real time traffic capabilities evaluation of a hybrid testbed for WiMAX networks with smart antenna support,” 2nd International Workshop on Multimode Wireless Access Networks, colocated with MOBILIGHT2011 conference, Bilbao, Spain, pp. 258–266, 9-11 May 2011.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Effective Frequency Plan Scheme for Downlink Coordinated Multi-point Transmission in LTE-A System

Xiaowei Liu
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: chelseamily@gmail.com

Xiaofeng Zhong
Dept. Electronic Engineering, Tsinghua National Laboratory for Information Sciences and Technology
Tsinghua University
Beijing, China
e-mail: zhongxf@tsinghua.edu.cn

***Abstract*—Coordinated multi-point (CoMP) transmission for LTE-Advanced is to improve system performance, especially to enhance cell coverage and cell-edge data rate. With efficient resource allocation schemes, CoMP transmission could be a promising technology to increase system throughput. Existing frequency plan schemes on downlink CoMP transmission limit the performance of the system since they do not take the changes in traffic load as well as link quality into consideration. In this paper, an optimized scheme is proposed that each base station could allocate different proportion of bandwidth to different users' equipments (UEs) according to their traffic load and link quality, to make the system more flexible and efficient. The scheme is more effective in both system throughput and cell-edge throughput as verified through system level simulation.**

Keywords-frequency plan; CoMP; LTE-A; system throughput; cell-edge throughput

I. INTRODUCTION

Coordinated multi-point transmission/reception is considered for LTE-Advanced as a tool to improve the coverage of high data rates, the cell-edge throughput and system throughput. Downlink coordinated multi-point transmission implies dynamic coordination among multiple geographically separated transmission points [1], that could improve cell-edge performance and to improve system throughput with efficient resource allocation [2][3]. In general, the cost of downlink CoMP for single user (SU) is found only beneficial to the cell-edge users where the perceived Signal-to-Interference-and-Noise Ratio (SINR) is low [4].

In order to improve system performance, several frequency plan schemes have been proposed. The fixed frequency plan scheme divides each cell's frequency into two parts, that is, CoMP Frequency Zone and Frequency Band for Single Cell Operation [5]. Such scheme maybe easy to implement, however, when considering the complexity of wireless communication environment, it lacks flexibility and may cause a decrease in system throughput performance. Another type of frequency plan scheme is the completely dynamic frequency plan scheme, known as the flexible frequency allocation plan (FFAP) scheme. Such plan ensures the cells which have formed more CoMP links to UEs being allocated more resource. The FFAP scheme allocates resource according to the number of cell-edge UEs, which makes it more flexible than the former fixed frequency plan scheme. However, such scheme allocates a certain proportion of frequency resource, which decreases system performance; and its complexity problem of joint schedule algorithms will become increasingly high [5].

Some of the recent works focuses on the FFAP scheme and its optimization [6]. However, since a certain proportion of frequency resource should be allocated to ensure CoMP transmission in such scheme, system performance may as well be decreased. Other recent works mainly focuses on schemes without frequency zone partition as well as algorithms to reduce implementation difficulties [7].

This paper proposes a frequency plan scheme to optimize the performance of the system considering both the traffic load as well as the link quality of UEs. Since it allocates frequency resource according to the cell's traffic load distribution and the link qualities, the proposed frequency plan scheme could achieve better fairness and system performance.

The remainder of the paper is organized as follows: Section II introduces the models of the system; Section III describes the proposed scheme for downlink SU-CoMP transmission. Simulation scenario and results are shown in Section IV. Section V concludes the paper.

II. SYSTEM MODEL WITH COMP TRANSMISSION

In CoMP transmission systems, cells could be classified into two different types: the serving cells and the coordinated cells. Each base station possess N_t transmit antennas while each UE consists of N_r receive antennas. The transmission model is described as follows.

A. Baseline Transmission

Baseline transmission refers to a situation of non-CoMP transmission. Under such system, each base station delivers data with individually selected codebook. Thus, the subordinate UE receive both data signal and interference from neighboring cells, with additive white Gaussian Noise. As in baseline transmission, there is no

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

cooperation performed between the serving cell and the neighboring cells. Under baseline transmission mode, considering N_t as transmit antennas and N_r as receive antennas, the received signal can be formulated as

$$r(i,j) = H(i,j)\cdot U_j \cdot s(i,j) + \sum_{\substack{m=1 \\ m\neq i}}^{M}\sum_{n=1}^{N} H(m,j)\cdot U_n \cdot s(m,n) + n(i,j). \quad (1)$$

where i and m are the indexes of base station, j and n are the indexes of UE, $r(i,j)$, $H(i,j)$, U_j, $s(i,j)$, $n(i,j)$ are the $N_r \times 1$ received signal, the $N_r \times N_t$ channel matrix, the $N_t \times 1$ precoding matrix, the transmit signal, and the $N_r \times 1$ additive white Gaussian noise (AWGN) matrix respectively [7]. Here, the first term on the right side of the equation denotes the power of desired signal ($s_{desired}$), the second denotes the power of interferences received (I), and the third denotes the power of noise received (N). The SINR under such mode could be calculated through the following equation

$$SINR = \frac{s_{desired}}{I + N}. \quad (2)$$

B. Downlink SU-CoMP Transmission

CoMP transmission consists of SU-CoMP mode and multiuser (MU) CoMP mode, where SU-CoMP refers to single user CoMP transmission and MU-CoMP refers to multiple user CoMP transmission [8]. Under SU-CoMP mode, the serving cell and the coordinated cell maintain the same codebook, which means SU-CoMP users can receive desired signals from both cells. Therefore, SU-CoMP users experience only interference from other cells (besides the serving and coordinated cells), with additive white Gaussian Noise.

In addition, the serving cells can both serve its subordinate UE as well as serve UE that belongs to other cells as coordinated cells. In most of the cases, cell-center UE would maintain better SINR, which leads to its often being served by their serving cells only. For cell-edge UE, their SINR may be so unpleasant that calls for coordinated cells to co-serve themselves. Therefore the definition of cell-center UE and cell-edge UE is very important in deciding the transmission scheme.

In our work, cell-edge UE could be defined as such UE, whose position is very near the boundary of its original serving cell, maintain a SINR that is relatively low. Such UE often calls for downlink CoMP transmission. Otherwise, UE whose position is in the center of its serving cell possesses a SINR that is relatively high denotes cell-center UE. Cell-center UE does not require coordinated cells to serve themselves; in other words, they have only one access point.

Under SU-CoMP transmission mode, the received signal could be formulated as:

$$r(i,j) = H(i,j)\cdot U_{p,j} \cdot s(i,j) + H(p,j)\cdot U_{p,j} \cdot s(p,j) + \sum_{\substack{m=1 \\ m\neq i,p}}^{M}\sum_{n=1}^{N} H(m,j)\cdot U_n \cdot s(m,n) + n(i,j). \quad (3)$$

where p is the index of coordinated cell base station [7]. Likewise, the first two terms on the right side of the equation denote signals received from the serving cell and the coordinated cell respectively, and combined as the desired signal $s_{desired}$. The third term denotes the interferences I , and the forth one denotes the noise received N . Therefore, the SINR under SU-CoMP transmission could be analyzed through (2).

When CoMP transmission is performed, resource is allocated according to frequency plan schemes. The fixed frequency plan scheme allocates a certain fixed portion of frequency band for CoMP transmission. Therefore, the frequency band of each cell is divided into two parts: the CoMP transmission frequency band and the serving frequency band, as shown in Fig. 1.

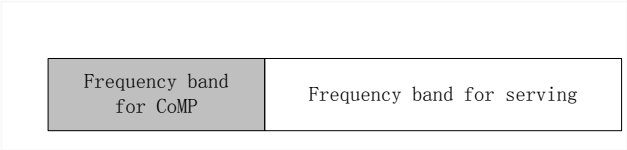


Figure 1. Fixed Frequency Plan Scheme

In such case, each cell retains bandwidth for CoMP transmission when CoMP happens. We could also estimate the best distribution method with good system performance. Such spectrum plan, though easy to operate, could hardly take into consideration the different situations between different cells. The FFAP scheme will have better utilization of the spectral bandwidth since it allocates frequency band according to the number of CoMP links in a certain cell. However, if CoMP Frequency Zone of one cell changes, other cells may also need adjustment to maintain CoMP operation. The FFAP scheme allocates a certain proportion of frequency bandwidth according to (4), which leads to the fact that is the same within each of the three cells.

$$\alpha = \frac{N_{cell-edge\,UE}}{N_{cell-center\,UE} + N_{cell-edge\,UE}}. \quad (4)$$

where $N_{cell-edge\,UE}$ and $N_{cell-center\,UE}$ represent the number of cell-edge UEs and the number of cell-center UEs in a certain cell, respectively.

III. The Proposed Frequency Allocation Plan Scheme

In our work, a new scheme is proposed to increase the flexibility and effectiveness of the system and system

performance under such plan could be greatly improved. The proposed scheme suggests that the proportion of frequency band allocates to SU-CoMP transmission is dynamic, that is to say, base stations could gather traffic load information of the system and link quality and therefore make decisions about the proportion could be chosen based on such knowledge.

The proposed frequency allocation plan scheme could be seen in Fig. 2, where BW is the total bandwidth of the spectrum.

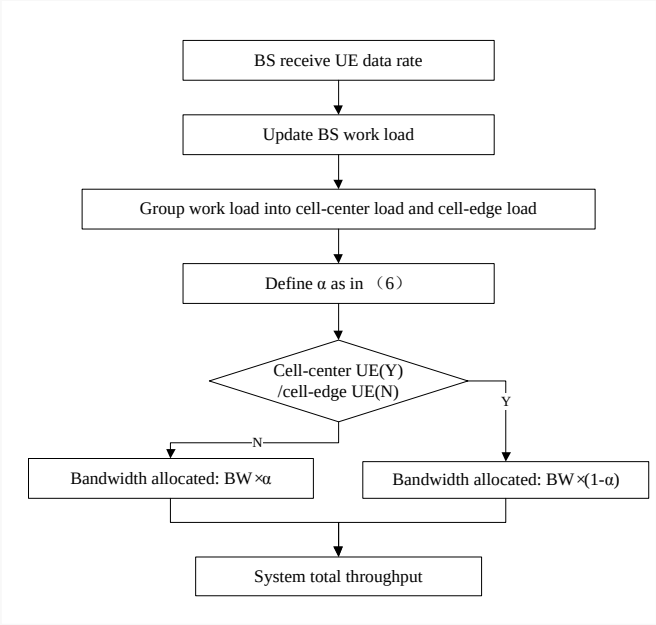


Figure 2. Flow Chart of Proposed Frequency Allocation Plan Scheme

On the first step, base stations receive information about data rate from both cell-center UEs and cell-edge UEs. In order to determine the access mode of a UE, that is, whether this user is served by a single serving base station or is served by both serving base station and coordinated serving station, a threshold ratio is defined. Different threshold ratio may result in different cell edge coverage. Therefore in our work the threshold ratio is decided according to the most likely percentage of CoMP users by simulation. In our work, the decision of CoMP transmission mode is operated by examining the following inequality among each UE.

$$\frac{RSRP_{neighbor}}{RSRP_{serving}} > threshold\ ratio. \qquad (5)$$

Here, RSRP refers to the Reference Signal Receiving Power. If (5) is true, we define the UE as cell-edge UE and it is served by both serving and coordinated cell, in other words, is under coordinated multipoint transmission mode. Otherwise, it is a cell-center UE and it is served by only its serving cell.

Then, base stations work out both the cell-center load and the cell-edge load based on the traffic load and link qualities of UEs. Such information could be acquired by base stations using channel information feedback. Therefore, the proportion of bandwidth assigned for CoMP users could be defined as:

$$\beta = \begin{cases} \dfrac{\log_2(\dfrac{1}{cell-edge\,load})}{\dfrac{1}{cell-center\,load}+\dfrac{1}{cell-edge\,load}}, & \text{cell-edge load} \neq 0, \\ 0, & \text{cell-edge load=0.} \end{cases} \qquad (6)$$

where $cell-edge\,load$ and $cell-center\,load$ are the total load of cell-edge UEs and the total load of cell-center UEs in a certain cell, respectively. Here, the use of function log is to ensure that β will not be overly high when $cell-edge\,load$ is relatively light. The algorithm is fairer than the FFAP plan since it allocates more resource to UEs that maintain relatively low link quality.

IV. SIMULATION RESULTS

TABLE I. SIMULATION PARAMETERS

Parameter	Assumption
Cellular Layout	Hexagonal grid, 3 sites,
Inter-site distance	100m
Load	Different number of UE per sector uniformly dropped
Subcarrier bandwidth	60kHz
Number of subcarriers per cell	128
Bandwidth	$60kHz \times 128 = 7.68MHz$
BS TX power per channel	-20dBm
Noise figure at UE	9dB
Lognormal Shadowing with shadowing standard deviation	8 dB
Distance path loss model	$PL(dB) = 34.5 + 35 \times log_{10} d$ d is in meters [9].
Traffic model	Random request
Channel Estimation	Ideal

A. Basic Scenario

We assume a three cell system with several UEs scenario to perform system level simulation.

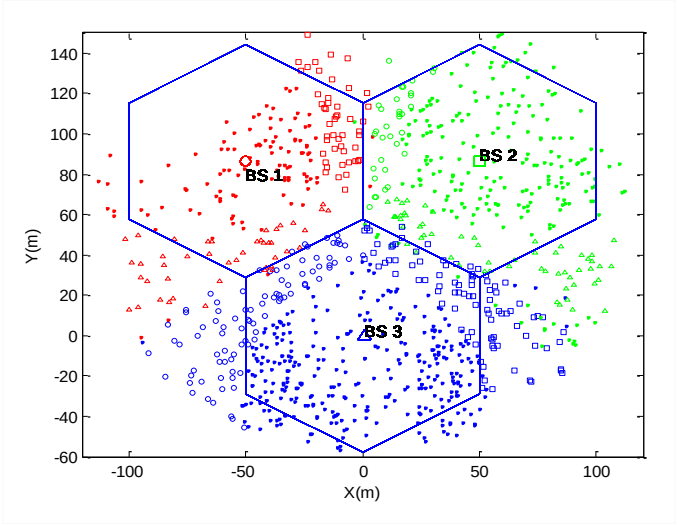


Figure 3. Location Map of Base Station and Mobile Pattern of Ues

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

As shown in Fig. 3, each marker denotes the location of UEs at a certain time. And the three base stations are marked as 'BS1', 'BS2', and 'BS3', with markers red ' ○ ', green ' □ ', and blue ' △ ', respectively. Each location point of UEs is stamped with a marker. The shape of the marker represents the coordinated cell of this UE, if exists, whereas the color of the marker represents the serving cell of this UE. If a user belongs to cell-center category, the shape of its marker will be a ' • '. For example, suppose UE_1 is served by cell 1 and is coordinated served by cell 2. The marker of UE_1 in the Location Map will be a red ' □ '.

B. Simulation Flow Chart

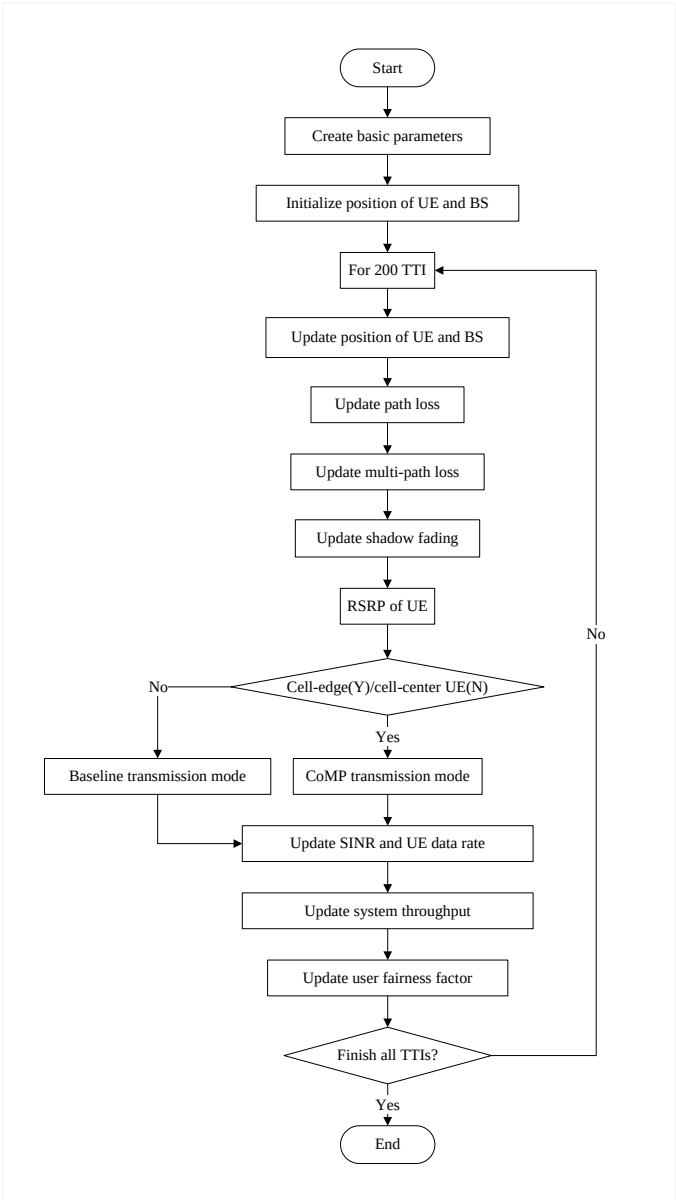


Figure 4. Flow Chart of System Level Simulation

Fig. 4 illustrates the steps of system level simulation of our work. The simulation applies 200 TTIs. First of all, in the step of initializing position of UE and BS, a scenario shown in Fig. 3 is set. Second, in every loop, path loss, multi-path loss, and shadow fading are updated. Then base stations are able to get information about channel matrix and therefore RSRP of UEs. After acquiring these qualities, the system group UEs into two categories, the cell-edge UEs and the cell-center UEs. Cell-edge UEs not only access to their serving cells, but also are assigned certain bandwidth to perform coordinated transmission, that is to say, is accessed by coordinated cells. Cell-center UEs, however, have only one access point, their serving cells, and exchange data through baseline transmission mode. Finally, the performance of the system under the proposed frequency allocation plan scheme is evaluated. SINR and user data rate are evaluated to obtain system throughput. Here, in order to keep the fairness of all UEs at a relatively high target, if the user data rate is below a threshold value, as:

$$Rate_{UE} < Rate_{threshold}. \quad (7)$$

This UE will be inactivated and rendered no spectrum resource at the moment, in other words, is disabled.

C. Simulation Results

Fig. 5 illustrates the total throughput of the system under three different frequency allocation plans, with the number of UE in three cells from 200 to 40,000. From the result of Fig. 5 we could conclude that the proposed scheme has a significant improvement in system throughput performance. For example, the gain of system throughput when performing the proposed scheme is 28.20% over the FFAP scheme (when the number of UE in the system is 1600).

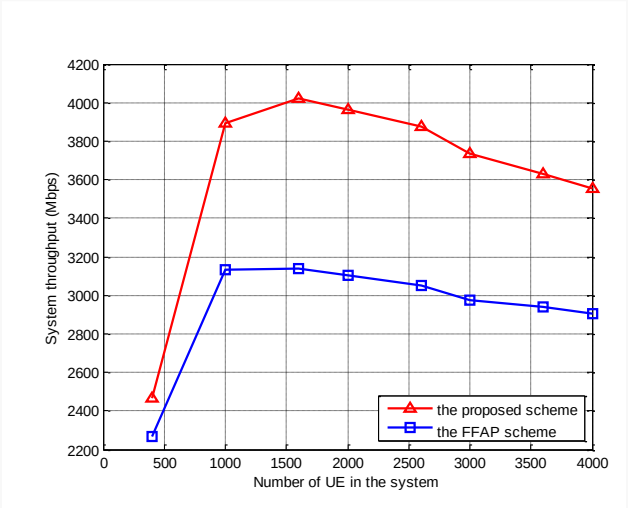


Figure 5. System Throughput Under Different Frequency Allocation Plan

Fig. 6 illustrates the system performance in cell-edge throughput, with the same condition as in Fig. 5. Cell-edge performance maintains an important indicator in SU-

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

CoMP system, which is also the main reason for downlink CoMP transmission. The proposed scheme shows higher gains over the FFAP scheme. 30.46% of the cell-edge throughput is improved when performing the proposed scheme. Therefore, cell-edge throughput performance should reach a preferred value when operating the proposed frequency allocation plan scheme.

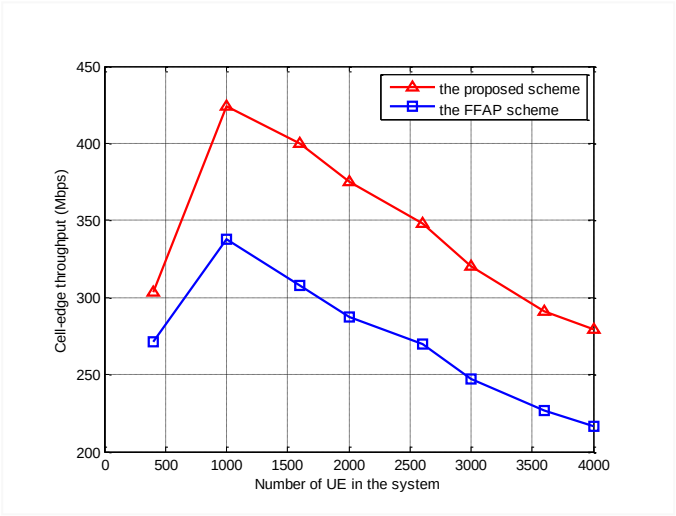


Figure 6. Cell-edge Throughput Under Different Frequency Allocation Plan

V. CONCLUSION

In this paper, we have evaluated the FFAP scheme in downlink CoMP transmission. A novel frequency allocation plan scheme is proposed, which takes both the traffic load of each cells and link quality into consideration. We have shown the effectiveness of the proposed scheme through system level simulation. About 30% gains of the system throughput and cell-edge throughput could be reached when performing the proposed scheme for resource allocation.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (60832008) and National S&T Major Project (2010ZX03003-002-02), and Chinese 973 Programme (2012CB316006). The work is finished by the visiting student in Tsinghua.

REFERENCES

[1] 3GPP R1-091263, CoMP Coordinated Scheduling for LTE-Advanced, March 2009.

[2] 3GPP R1-090793, Coordinated Multi-Point Transmission — Coordinated Beamforming and Results, February 2009.

[3] Batista, R.L., dos Santos, R.B., Maciel, T.F., Freitas, W.C., Cavalcanti, F.R.P., Performance Evaluation for Resource Allocation Algorithms in CoMP Systems, Proceedings of the 2010 IEEE 72nd Vehicular Technology Conference, IEEE Press, 6-9 Sept. 2010, pp. 1-5, doi:10.1109/VETECF.2010.5594241.

[4] Akyildiz, I. F., Gutierrez-Estevez, D. M. and Chavarria-Reyes, E., The evolution to 4G cellular systems: LTE-Advanced, Physical Communications (Elsevier) Journal, vol. 3, no. 4, pp. 217-244, December 2010.

[5] 3GPP R1-091415, Further Discussion of Frequency Plan scheme on CoMP-SU-MIMO, March 2009.

[6] Xingkun Xu, Tao Qiu , Wenjun Xu, Zhiqiang He and Kai Niu, Subcarrier allocation combined with coordinated multi-point transmission in multi-cell OFDMA system, Proceedings of the 2009 IEEE International Conference on Network Infrastructure and Digital Content, IEEE Press, 6-8 Nov. 2009, pp. 842-846, doi:10.1109/ICNIDC.2009.5360813.

[7] Jing LIU, Yongyu CHANG, Qun PAN, Xin ZHANG, and Dacheng YANG, A novel transmission scheme and scheduling algorithm for CoMP-SU-MIMO in LTE-A system, Proceedings of the 2010 IEEE 71st Vehicular Technology Conference, IEEE Press, 16-19 May 2010, pp. 1-5, doi:10.1109/VETECS.2010.5493812.

[8] 3GPP R1-094252, Different Types of DL CoMP Transmission for LTE-A, October 2009.

[9] 3GPP TR 25.996 V8.0.0, Spatial channel model for Multiple Input Multiple Output (MIMO) simulations, Dec. 2008.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Call Blocking Probabilities of Elastic and Adaptive Traffic with Retrials

Ioannis D. Moscholios*, Vassilios G. Vassilakis†, John S. Vardakas† and Michael D. Logothetis†
*Dept. of Telecommunications Science and Technology, University of Peloponnese, 221 00 Tripolis, Greece.
Email: idm@uop.gr
†Dept. of Electrical and Computer Engineering, University of Patras, 265 04 Patras, Greece.
{vasilak, jvardakas, mlogo}@upatras.gr

***Abstract*—We study a single-link multirate loss system, which accommodates both elastic and adaptive traffic of Poisson arriving calls, with exponentially distributed service time and flexible bandwidth requirements. If the available link bandwidth is not enough to accept a new call with its peak-bandwidth requirement, then the call can retry one or more times (single/multi-retry loss models) to be connected in the system with reduced bandwidth. If its last bandwidth requirement remains higher than the available link bandwidth, the call can still be accepted in the system, by compressing the bandwidth of all in-service calls together with its last bandwidth requirement. The proposed models do not have a product form solution, and therefore we propose approximate recursive formulas for the efficient calculation of call blocking probabilities. The consistency and the accuracy of our models are verified by simulation, and found to be very satisfactory.**

***Keywords*-Poisson process; elastic/adaptive traffic; call blocking; Markov chains; recurrent formula.**

I. INTRODUCTION

The call-level QoS assessment in modern telecom networks remains an open issue, due to the existence of elastic and adaptive traffic. By the term "elastic traffic" we mean calls that can compress their bandwidth, while simultaneously increasing their service time, during their lifetime in the system, so that the product *service time* by *bandwidth* remains constant. In the case of "adaptive traffic", calls can compress their bandwidth but they do not alter their service time. The call-level modeling of elastic and adaptive traffic is mostly based on the classical Erlang Multirate Loss Model (EMLM) ([[1]- [2]) which has been widely used in wired (e.g. [3]- [5]), wireless (e.g. [6]- [7]) and optical networks (e.g. [8]- [10]) to model systems that accommodate calls of different service-classes.

In the EMLM, Poisson arriving calls of different service-classes compete for the available link bandwidth under the complete sharing policy (all calls compete for all bandwidth resources). Calls are blocked and lost if their required bandwidth is higher than the available link bandwidth. Accepted calls remain in the link for a generally distributed service time [1]. The fact that the steady-state probabilities of the EMLM have a Product Form Solution (PFS) leads to an accurate calculation of Call Blocking Probabilities (CBP) via the Kaufman-Roberts recursive formula [1], [2]. In [11], the EMLM is extended to include retrials. Blocked calls can immediately retry one or more times (Single- and Multi-Retry loss Model, SRM and MRM, respectively) to be connected in the system by requiring less bandwidth units (b.u.). A retry call is blocked and lost if its last bandwidth requirement is higher than the available link bandwidth.

In this paper, we consider a system supporting elastic and adaptive traffic with single/multi retrials. If the available link bandwidth is less than or equal to the last bandwidth requirement of a retry call, the system compresses it, down to a minimum proportion (common to all calls) of its required last bandwidth, together with the bandwidth of all in-service calls. If the resulting bandwidth is less than the available link bandwidth, the new call is accepted; otherwise is blocked and lost. When a call, whose bandwidth is compressed, departs from the system, then all in-service calls expand their bandwidth. Due to retrials/compression, the models (single and multi-retry model) do not have a PFS. However, we propose approximate recursive formulas for the link occupancy distribution, and consequently CBP, calculation. Simulation results validate the proposed models and show very good accuracy. If only elastic traffic exists in the link, then the proposed models coincide with the models of [12]. In the case of no retrials for calls of all service-classes, the proposed models coincide with the model of [13], which is named, herein, Extended EMLM (E-EMLM). In [14], elastic/adaptive calls have several bandwidth requirements and request for bandwidth, upon their arrival, according to the occupied link bandwidth (i.e. calls do not retry).

This paper is organized as follows. In Section II, we review the E-EMLM, the SRM and the MRM. In Section III, we present the proposed models and provide formulas for the approximate calculation of CBP. In Section IV, we present analytical and simulation results to evaluate the accuracy and consistency of our models. We conclude in Section V.

II. REVIEW OF THE E-EMLM AND MULTIRATE LOSS MODELS WITH RETRIALS

A. Review of the E-EMLM

Consider a link of capacity C b.u. that accommodates K service-classes and let $T>C$ be the limit that determines the maximum permitted bandwidth compression among calls; the higher parameter T, the higher permitted compression. A service-class k ($k = 1, \ldots, K$) can be elastic or adaptive.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Let K_e and K_a be the set of elastic and adaptive service-classes ($K_e+K_a=K$), respectively. Service-class k calls follow a Poisson process with rate λ_k, request b_k b.u. (peak-bandwidth requirement) and have an exponentially distributed service time with mean μ_k^{-1}. Let j be the occupied link bandwidth, $j=0,1,\ldots,T$, when a service-class k call arrives in the link. If $j+b_k \le C$, the call is accepted in the link with b_k, μ_k^{-1}. If $j+b_k > T$ the call is blocked and lost. If $T \ge j+b_k > C$ the call is accepted in the link by compressing its bandwidth and the bandwidth of all in-service calls. The compressed bandwidth of the new service-class k call is given by:

$$b_k^{'} = rb_k = \frac{C}{j^{'}} b_k \tag{1}$$

where r is the compression factor (common to all service-classes) given by $r \equiv r(\boldsymbol{n}) = C/j$, $j^{'} = j + b_k = \boldsymbol{nb} + b_k$, $\boldsymbol{n} = (n_1, n_2, ..., n_k, ..., n_K)$, $\boldsymbol{b} = (b_1, b_2, ..., b_k, ..., b_K)$ and n_k is the number of in-service calls of service-class k in steady state.

Similarly, all in-service calls compress their bandwidth to $b_i^{'} = \frac{C}{j^{'}} b_i$ for $i = 1, ..., K$. After the compression of all calls the link state is $j = C$. So, the link operates at its full capacity and all calls share this capacity in proportion to their peak-bandwidth requirement. The minimum bandwidth that a service-class k call (new or in-service) tolerates is:

$$b_{k,\min}^{'} = r_{\min} b_k = \frac{C}{T} b_k \tag{2}$$

where $r_{\min}$ is the minimum proportion of peak-bandwidth.

When a service-class k call, with bandwidth $b_k^{'}$, departs from the system, the remaining in-service calls of each service-class i $(i = 1, ..., K)$, expand their bandwidth in proportion to their peak-bandwidth b_i. After bandwidth compression/expansion, all elastic service-class k calls ($k = 1, ..., K_e$) increase/decrease their service time so that the product *service time* by *bandwidth* remains constant. Adaptive service-class calls do not alter their service time.

The bandwidth compression mechanism destroys reversibility in the model and therefore no PFS exists. However, in [13] an approximate recursive formula is proposed for the calculation of the link occupancy distribution, $G(j)$:

$$G(j) = \begin{cases} 1 & \text{for } j = 0 \\ \frac{1}{\min(j,C)} \sum\limits_{k \in K_e} \alpha_k b_k G(j - b_k) + \frac{1}{j} \sum\limits_{k \in K_a} \alpha_k b_k G(j - b_k) & \\ & \text{for } j = 1, ..., T \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $\alpha_k = \lambda_k \mu_k^{-1}$ is the offered traffic-load (in erl) of service-class k calls.

The proof of (3) is based on a reversible Markov chain which approximates the bandwidth compression/expansion mechanism of the E-EMLM. The local balance (LB) equations of this Markov chain are of the form [13]:

$$\lambda_k P(\boldsymbol{n}_k^-) = n_k \mu_k \phi_k(\boldsymbol{n}) P(\boldsymbol{n}) \tag{4}$$

where $P(\boldsymbol{n})$ is the probability distribution of state $\boldsymbol{n}$, $P(\boldsymbol{n}_k^-)$ is the probability distribution of $\boldsymbol{n}_k^- = (n_1, n_2, ..., n_{k-1}, n_k - 1, n_{k+1}, ..., n_K)$ and $\phi_k(\boldsymbol{n})$ is a state dependent factor which describes: i) the bandwidth compression factor and ii) the increase factor of service time of service-class k $(k = 1, ..., K)$ calls in state $\boldsymbol{n}$. In other words, $\phi_k(\boldsymbol{n})$ has the same role with r in (1) but it may be different for each service-class. The values of $\phi_k(\boldsymbol{n})$ are determined by:

$$\phi_k(\boldsymbol{n}) = \begin{cases} 1 & , \text{ for } \boldsymbol{nb} \le C, \ \boldsymbol{n} \in \boldsymbol{\Omega} \\ \frac{x(\boldsymbol{n}_k^-)}{x(\boldsymbol{n})} & , \text{ for } C < \boldsymbol{nb} \le T \ \boldsymbol{n} \in \boldsymbol{\Omega} \\ 0 & , \text{ otherwise} \end{cases} \tag{5}$$

where $\boldsymbol{\Omega} = \{\boldsymbol{n} : 0 \le \boldsymbol{nb} \le T\}$ and $\boldsymbol{nb} = \sum_{k=1}^{K} n_k b_k$.

In (5), $x(\boldsymbol{n})$ is a state multiplier, whose values are chosen so that (4) holds, [13]:

$$x(\boldsymbol{n}) = \begin{cases} 1 & , \text{ when } \boldsymbol{nb} \le C, \ \boldsymbol{n} \in \boldsymbol{\Omega} \\ \frac{1}{C} \left(\sum\limits_{k \in K_e} n_k b_k x(\boldsymbol{n}_k^{-1}) + r(\boldsymbol{n}) \sum\limits_{k \in K_a} n_k b_k x(\boldsymbol{n}_k^{-1}) \right) & \\ \qquad \text{when } C < \boldsymbol{nb} \le T, \ \boldsymbol{n} \in \boldsymbol{\Omega} & \\ 0 & , \text{ otherwise} \end{cases} \tag{6}$$

Having determined $G(j)$ we calculate CBP of service-class k calls, B_k, as follows:

$$B_k = \sum_{j=T-b_k+1}^{T} G^{-1} G(j) \tag{7}$$

where $G = \sum_{j=0}^{T} G(j)$ is the normalization constant.

B. Review of multirate loss models with retrials

Consider again a link of capacity C b.u. that accommodates Poisson arriving calls of K service-classes. Calls of service-class k $(k = 1, ..., K)$ have an arrival rate λ_k and request b_k b.u. If b_k b.u. are available, a call of service-class k remains in the system for an exponentially distributed service-time with mean μ_k^{-1}. Otherwise, the call is blocked and retries to be connected with parameters (b_{kr}, μ_{kr}^{-1}) where $b_{kr} < b_k$ and $\mu_{kr}^{-1} > \mu_k^{-1}$. The SRM does not have a PFS and therefore the determination of $G(j)$, is based on an approximate recursive formula, [11]:

$$G(j) = \begin{cases} 1 & , \text{ for } j = 0 \\ \frac{1}{j} \sum\limits_{k=1}^{K} \alpha_k b_k G(j - b_k) + & \\ \frac{1}{j} \sum\limits_{k=1}^{K} \alpha_{kr} b_{kr} \gamma_{kr}(j) G(j - b_{kr}) & \text{for } j = 1, ..., C \\ 0 & , \text{ otherwise} \end{cases} \tag{8}$$

where: $\alpha_k = \lambda_k \mu_k^{-1}$, $\alpha_{kr} = \lambda_k \mu_{kr}^{-1}$, $\gamma_{kr}(j) = 1$ when $j > C - (b_k - b_{kr})$.

The proof of (8) is based on two assumptions: 1) the application of LB, which exists only in PFS models and 2) the application of *Migration Approximation* (MA) which assumes that the occupied link bandwidth from retry calls

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

is negligible when $j \leq C-(b_k-b_{kr})$. The variable $\gamma_{kr}(j)$ expresses the MA in (8). The blocking probability of a retry service-class k call, B_{kr}, is given by:

$$B_{kr} = \sum_{j=C-b_{kr}+1}^{C} G^{-1}G(j) \quad (9)$$

where $G=\sum_{j=0}^{C} G(j)$ is the normalization constant.

In the MRM, a blocked service-class *k* call may retry many times with parameters $(b_{kr_l}, \mu_{kr_l}^{-1})$ for $l=1,...,s(k)$, where $b_{kr_{s(k)}}<...<b_{kr_1}<b_k$ and $\mu_{kr_{s(k)}}^{-1}>...>\mu_{kr_1}^{-1}>\mu_k^{-1}$. The MRM does not have a PFS and therefore the calculation of *G(j)*, is based on an approximate recursive formula [11]:

$$G(j)=\begin{cases} 1 & \text{for } j=0 \\ \frac{1}{j}\sum_{k=1}^{K}\alpha_k b_k G(j-b_k)+ & \\ \frac{1}{j}\sum_{k=1}^{K}\sum_{l=1}^{s(k)}\alpha_{kr_l} b_{kr_l}\gamma_{kr_l}(j)G(j-b_{kr_l}) & \text{for } j=1,...,C \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where: $\alpha_{kr_l}=\lambda_k\mu_{kr_l}^{-1}$ and $\gamma_{kr_l}(j)=1$, if $C\geq j>C-(b_{kr_{l-1}}-b_{kr_l})$.

The blocking probability of a retry service-class *k* call with its last bandwidth requirement, $B_{kr_{s(k)}}$, is given by:

$$B_{kr_{s(k)}} = \sum_{j=C-b_{kr_{s(k)}}+1}^{C} G^{-1}G(j) \quad (11)$$

If calls of a service-class *k* do not have retry parameters, then their blocking probability, B_k, is determined by:

$$B_k = \sum_{j=C-b_k+1}^{C} G^{-1}G(j) \quad (12)$$

III. Multirate Loss Models of Elastic & Adaptive Traffic With Retrials

A. The elastic-adaptive single-retry loss model

The proposed Elastic-Adaptive Single-Retry loss Model (EA-SRM) is a non-PFS model that combines the characteristics of the E-EMLM and the SRM. In order to prove an approximate but recursive formula for the calculation of *G(j)*, the following example is presented.

Consider a link of capacity *C* b.u. that accommodates Poisson arriving calls of two service-classes. The 1^{st} service-class is adaptive and the 2^{nd} is elastic. Only calls of the 2^{nd} service-class have retry parameters. The traffic parameters are: $(\lambda_1, \mu_1^{-1}, b_1)$ and $(\lambda_2, \mu_2^{-1}, \mu_{2r}^{-1}, b_2, b_{2r})$ with $b_{2r}<b_2$ and $\mu_{2r}^{-1}>\mu_2^{-1}$. Bandwidth compression is permitted for calls of both service-classes up to a limit *T*. Although the EA-SRM is a non-PFS model we use the LB of (4), initially for calls of the 1^{st} service-class:

$$\lambda_1 P(\boldsymbol{n}_1^-) = n_1\mu_1\phi_1(\boldsymbol{n})P(\boldsymbol{n}), \quad 1 \leq \boldsymbol{nb} \leq T \quad (13)$$

where $\boldsymbol{n}=(n_1,n_2,n_{2r}), \boldsymbol{n}_1^-=(n_1-1,n_2,n_{2r})$ with $n_1\geq 1$ and

$$\phi_1(\boldsymbol{n}) = \begin{cases} 1 & , \text{ for } \boldsymbol{nb}\leq C, \ \boldsymbol{n}\in\boldsymbol{\Omega} \\ \frac{x(\boldsymbol{n}_1^-)}{x(\boldsymbol{n})}, & \text{for } C<\boldsymbol{nb}\leq T, \ \boldsymbol{n}\in\boldsymbol{\Omega} \\ 0 & , \text{ otherwise} \end{cases} \quad (14)$$

with $\boldsymbol{nb} = j = n_1b_1+n_2b_2+n_{2r}b_{2r}$.

Based on (14) and by multiplying both sides of (13) with b_1 and $r(\boldsymbol{n})$ we have:

$$\alpha_1 b_1 x(\boldsymbol{n})r(\boldsymbol{n})P(\boldsymbol{n}_1^-)=n_1b_1x(\boldsymbol{n}_1^-)r(\boldsymbol{n})P(\boldsymbol{n}), \ 1\leq\boldsymbol{nb}\leq T \quad (15)$$

where $\alpha_1=\lambda_1\mu_1^{-1}$ and $r(\boldsymbol{n})=\min(1,C/j)$.

The LB equations of the 2^{nd} service-class calls are derived as follows:

a) If a call arrives in the system and $j+b_2\leq C$ then it is accepted with b_2 b.u. Thus, the following LB equation holds:

$$\lambda_2 P(\boldsymbol{n}_2^-)=n_2\mu_2\phi_2(\boldsymbol{n})P(\boldsymbol{n}), \quad 1\leq\boldsymbol{nb}\leq C \quad (16)$$

where $\phi_2(\boldsymbol{n}) = \frac{x(\boldsymbol{n}_2^-)}{x(\boldsymbol{n})}=1$, when $1\leq\boldsymbol{nb}\leq C$ and $\boldsymbol{n}_2^- = (n_1, n_2-1, n_{2r})$ with $n_2\geq 1$.

Multiplying both sides of (16) with b_2 we have:

$$\alpha_2 b_2 x(\boldsymbol{n})P(\boldsymbol{n}_2^-)=n_2b_2x(\boldsymbol{n}_2^-)P(\boldsymbol{n}), \ 1\leq\boldsymbol{nb}\leq C \quad (17)$$

where $\alpha_2=\lambda_2\mu_2^{-1}$.

b) If a call arrives in the system and $j+b_2>C$ then the call is blocked with b_2 and immediately retries with $b_{2r}<b_2$. Now if: 1) $j+b_{2r}\leq C$ the retry call is accepted in the system with b_{2r}, 2) $j+b_{2r}>T$ the retry call is blocked and lost and 3) $C<j+b_{2r}\leq T$ the retry call is accepted in the system by compressing its bandwidth requirement b_{2r} together with the bandwidth of all in-service calls. The compressed bandwidth of the retry call is $b_{2r}^{'}=rb_{2r}=\frac{C}{j+b_{2r}}b_{2r}$. Thus,

$$\lambda_2 P(\boldsymbol{n}_{2r}^-)=n_{2r}\mu_{2r}\phi_{2r}(\boldsymbol{n})P(\boldsymbol{n}), \text{ for } C-b_2+b_{2r}<\boldsymbol{nb}\leq T \quad (18)$$

where $P(\boldsymbol{n}_{2r}^-)$ is the probability distribution of state $\boldsymbol{n}_{2r}^- = (n_1, n_2, n_{2r}-1)$ with $n_{2r}\geq 1$ and

$$\phi_{2r}(\boldsymbol{n}) = \begin{cases} 1 & , \text{ for } \boldsymbol{nb}\leq C, \ \boldsymbol{n}\in\boldsymbol{\Omega} \\ \frac{x(\boldsymbol{n}_{2r}^-)}{x(\boldsymbol{n})}, & \text{for } C<\boldsymbol{nb}\leq T, \ \boldsymbol{n}\in\boldsymbol{\Omega} \\ 0 & , \text{ otherwise} \end{cases} \quad (19)$$

Based on (19) and by multiplying both sides of (18) with b_{2r} we have:

$$\alpha_{2r} b_{2r} x(\boldsymbol{n})P(\boldsymbol{n}_{2r}^-)=n_{2r}b_{2r}x(\boldsymbol{n}_{2r}^-)P(\boldsymbol{n}), \text{ for } C-b_2+b_{2r}<\boldsymbol{nb}\leq T \quad (20)$$

where $\alpha_{2r}=\lambda_{2r}\mu_{2r}^{-1}$.

Eqs. (15), (17) and (20) lead to a system of equations:

$$\begin{aligned} &\alpha_1 b_1 x(\boldsymbol{n})r(\boldsymbol{n})P(\boldsymbol{n}_1^-) + \alpha_2 b_2 x(\boldsymbol{n})P(\boldsymbol{n}_2^-) \\ &= (n_1 b_1 x(\boldsymbol{n}_1^-)r(\boldsymbol{n}) + n_2 b_2 x(\boldsymbol{n}_2^-))P(\boldsymbol{n}) \end{aligned} \quad (21)$$

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

for $1 \le \boldsymbol{nb} \le C - b_2 + b_{2r}$.

$$\begin{aligned}&\alpha_1 b_1 x(\boldsymbol{n}) r(\boldsymbol{n}) P(\boldsymbol{n}_1^-) + \alpha_2 b_2 x(\boldsymbol{n}) P(\boldsymbol{n}_2^-) + \alpha_{2r} b_{2r} x(\boldsymbol{n}) P(\boldsymbol{n}_{2r}^-) \\ &= (n_1 b_1 x(\boldsymbol{n}_1^-) r(\boldsymbol{n}) + n_2 b_2 x(\boldsymbol{n}_2^-) + n_{2r} b_{2r} x(\boldsymbol{n}_{2r}^-)) P(\boldsymbol{n})\end{aligned} \tag{22}$$

for $C - b_2 + b_{2r} < \boldsymbol{nb} \le C$.

$$\begin{aligned}&\alpha_1 b_1 x(\boldsymbol{n}) r(\boldsymbol{n}) P(\boldsymbol{n}_1^-) + a_{2r} b_{2r} x(\boldsymbol{n}) P(\boldsymbol{n}_{2r}^-) \\ &= (n_1 b_1 x(\boldsymbol{n}_1^-) r(\boldsymbol{n}) + n_{2r} b_{2r} x(\boldsymbol{n}_{2r}^-)) P(\boldsymbol{n})\end{aligned} \tag{23}$$

for $C < \boldsymbol{nb} \le T$.

Eqs. (21)-(23) are combined in one equation by assuming that calls with b_{2r} are negligible when $1 \le \boldsymbol{nb} \le C - b_2 + b_{2r}$ (MA) and calls with b_2 are negligible when $C < \boldsymbol{nb} \le T$:

$$\begin{aligned}&\alpha_1 b_1 x(\boldsymbol{n}) r(\boldsymbol{n}) P(\boldsymbol{n}_1^-) + \alpha_2 b_2 \gamma_2(\boldsymbol{nb}) x(\boldsymbol{n}) P(\boldsymbol{n}_2^-) + \\ &+ \alpha_{2r} b_{2r} \gamma_{2r}(\boldsymbol{nb}) x(\boldsymbol{n}) P(\boldsymbol{n}_{2r}^-) = \\ &(n_1 b_1 x(\boldsymbol{n}_1^-) r(\boldsymbol{n}) + n_2 b_2 x(\boldsymbol{n}_2^-) + n_{2r} b_{2r} x(\boldsymbol{n}_{2r}^-)) P(\boldsymbol{n})\end{aligned} \tag{24}$$

for $0 \le \boldsymbol{n} \le T$,
where $\gamma_2(\boldsymbol{nb}) = 1$ for $1 \le \boldsymbol{nb} \le C$, otherwise $\gamma_2(\boldsymbol{nb}) = 0$ and $\gamma_{2r}(\boldsymbol{nb}) = 1$ for $C - b_2 + b_{2r} < \boldsymbol{nb} \le T$, otherwise $\gamma_{2r}(\boldsymbol{nb}) = 0$.

At this point, we derive a formula for $x(\boldsymbol{n})$ by making the following assumptions:

1) When $C < \boldsymbol{nb} \le T$, $\boldsymbol{n} \in \boldsymbol{\Omega}$, the bandwidth of all in-service calls should be compressed by $\phi_k(\boldsymbol{n})$, $k = 1, 2$, so that:
$$n_1 b_1^{'} + n_2 b_2^{'} + n_{2r} b_{2r}^{'} = C \tag{25}$$

2) We keep the product *service time* by *bandwidth* of service-class k calls (elastic or adaptive) in state $\boldsymbol{n}$ of the irreversible Markov chain equal to the corresponding product in the same state $\boldsymbol{n}$ of the reversible Markov chain:
$$\begin{aligned}&\tfrac{b_1 r(\boldsymbol{n})}{\mu_1} = \tfrac{b_1^{'}}{\mu_1 \phi_1(\boldsymbol{n})} \quad \text{or } b_1^{'} = b_1 \phi_1(\boldsymbol{n}) r(\boldsymbol{n}) \\ &\tfrac{b_2 r(\boldsymbol{n})}{\mu_2 r(\boldsymbol{n})} = \tfrac{b_2^{'}}{\mu_2 \phi_2(\boldsymbol{n})} \quad \text{or } b_2^{'} = b_2 \phi_2(\boldsymbol{n}) \\ &\tfrac{b_{2r} r(\boldsymbol{n})}{\mu_{2r} r(\boldsymbol{n})} = \tfrac{b_{2r}^{'}}{\mu_{2r} \phi_{2r}(\boldsymbol{n})} \quad \text{or } b_{2r}^{'} = b_{2r} \phi_{2r}(\boldsymbol{n})\end{aligned} \tag{26}$$

By substituting (26) in (25) we obtain:

$$n_1 b_1 \phi_1(\boldsymbol{n}) r(\boldsymbol{n}) + n_2 b_2 \phi_2(\boldsymbol{n}) + n_{2r} b_{2r} \phi_{2r}(\boldsymbol{n}) = C \tag{27}$$

where $\phi_1(\boldsymbol{n})$, $\phi_2(\boldsymbol{n})$ are given by (14) and $\phi_{2r}(\boldsymbol{n})$ by (19).

Eq. (27), due to (14) and (19), is written as:

$$x(n) = \begin{cases} 1 & , \text{ for } \boldsymbol{nb} \le C, \ \boldsymbol{n} \in \boldsymbol{\Omega} \\ \frac{1}{C}\left[n_1 b_1 x(\boldsymbol{n}_1^-) r(\boldsymbol{n}) + n_2 b_2 x(\boldsymbol{n}_2^-) + n_{2r} b_{2r} x(\boldsymbol{n}_{2r}^-)\right] & \\ \qquad \text{for } C < \boldsymbol{nb} \le T, \ \boldsymbol{n} \in \boldsymbol{\Omega} & \\ 0 & , \text{ otherwise} \end{cases} \tag{28}$$

Based on (28), we consider again (24). Since $x(\boldsymbol{n}) = 1$, when $0 \le \boldsymbol{nb} \le C$, it is proved in [11] that:

$$\alpha_1 b_1 G(j - b_1) + \alpha_2 b_2 G(j - b_2) + \alpha_{2r} b_{2r} \gamma_{2r}(j) G(j - b_{2r}) = j G(j) \tag{29}$$

for $1 \le j \le C$ and $\gamma_{2r}(j) = 1$ for $C - b_2 + b_{2r} < j$.

To prove (29), the MA is needed, which assumes that 2nd service-class retry calls do not exist in states $j \le C - b_2 + b_{2r}$.

When $C < \boldsymbol{nb} \le T$, we have $\gamma_2(\boldsymbol{nb}) = 0$ and based on (28), we can write (24) as follows:

$$\frac{C}{j} \alpha_1 b_1 P(\boldsymbol{n}_1^-) + \alpha_{2r} b_{2r} \gamma_{2r}(\boldsymbol{nb}) P(\boldsymbol{n}_{2r}^-) = C P(\boldsymbol{n}) \tag{30}$$

since $r(\boldsymbol{n}) = C/j$, when $C < \boldsymbol{nb} \le T$.

To introduce the link occupancy distribution $G(j)$ in (30), we sum both sides of (30) over $\boldsymbol{\Omega}_j = \{\boldsymbol{n} \in \boldsymbol{\Omega} \,|\boldsymbol{nb} = j\}$:

$$\frac{C}{j} \alpha_1 b_1 \sum_{\{\boldsymbol{n}|\boldsymbol{nb}=j\}} P(\boldsymbol{n}_1^-) + \alpha_{2r} b_{2r} \gamma_{2r}(j) \sum_{\{\boldsymbol{n}|\boldsymbol{nb}=j\}} P(\boldsymbol{n}_{2r}^-) = C \sum_{\{\boldsymbol{n}|\boldsymbol{nb}=j\}} P(\boldsymbol{n}) \tag{31}$$

Since $\sum_{\boldsymbol{n} \in \boldsymbol{\Omega}_j} P(\boldsymbol{n}) = G(j)$, (31) is written as:

$$\frac{C}{j} \alpha_1 b_1 G(j - b_1) + \alpha_{2r} b_{2r} \gamma_{2r}(j) G(j - b_{2r}) = C G(j) \tag{32}$$

where $\gamma_{2r}(j) = 1$ for $C - b_2 + b_{2r} < j \le T$.

The combination of (29) and (32) gives an approximate recursive formula for the calculation of $G(j)$ (for $1 \le j \le T$) when the 1^{st} service-class is adaptive and the 2^{nd} service-class is elastic with retrials:

$$\begin{aligned}&G(j) = \tfrac{1}{j} \alpha_1 b_1 G(j - b_1) + \\ &\tfrac{1}{\min(j, C)} [\alpha_2 b_2 \gamma_2(j) G(j - b_2) + \alpha_{2r} b_{2r} \gamma_{2r}(j) G(j - b_{2r})]\end{aligned} \tag{33}$$

where $\gamma_2(j) = 1$ for $1 \le j \le C$, $\gamma_{2r}(j) = 1$ for $C - b_2 + b_{2r} < j \le T$.

In the case of K service-classes and assuming that all service-classes may have retry parameters, (33) becomes:

$$G(j) = \begin{cases} 1 & \text{for } j = 0 \\ \frac{1}{j}\left[\sum\limits_{k \in K_a} \alpha_k b_k \gamma_k(j) G(j - b_k) + \sum\limits_{k \in K_a} \alpha_{kr} b_{kr} \gamma_{kr}(j) G(j - b_{kr})\right] + & \\ \frac{1}{\min(C, j)}\left[\sum\limits_{k \in K_e} \alpha_k b_k \gamma_k(j) G(j - b_k) + \sum\limits_{k \in K_e} \alpha_{kr} b_{kr} \gamma_{kr}(j) G(j - b_{kr})\right] & \\ \qquad \text{for } j = 1, ..., T & \\ 0 \quad \text{otherwise} & \end{cases} \tag{34}$$

where: $\alpha_k = \lambda_k \mu_k^{-1}$, $\alpha_{kr} = \lambda_k \mu_{kr}^{-1}$,

$$\gamma_k(j) = \begin{cases} 1 \text{ for } 1 \le j \le C \text{ and } b_{kr} > 0 \\ 1 \text{ for } 1 \le j \le T \text{ and } b_{kr} = 0 \\ 0 \text{ otherwise} \end{cases}$$

$$\gamma_{kr}(j) = \begin{cases} 1 \text{ for } C - b_k + b_{kr} < j \le T \\ 0 \text{ otherwise} \end{cases}$$

The blocking probability of a retry service-class k call, B_{kr}, is given by:

$$B_{kr} = \sum_{j=T-b_{kr}+1}^{T} G^{-1} G(j) \tag{35}$$

where $G = \sum_{j=0}^{T} G(j)$ is the normalization constant.

B. The elastic-adaptive multi-retry loss model

In the proposed Elastic-Adaptive Multi-Retry loss Model (EA-MRM) a blocked service-class k call may have many retry parameters $(b_{kr_l}, \mu_{kr_l}^{-1})$ for $l = 1, ..., s(k)$, with $b_{kr_{s(k)}} <$

$... < b_k$ and $\mu^{-1}_{kr_{s(k)}} > ... > \mu^{-1}_k$. The EA-MRM does not have a PFS and therefore the calculation of $G(j)$ is based on an approximate formula whose proof is similar to that of (34):

$$G(j)=\begin{cases} 1 \qquad \text{for } j=0 \\ \frac{1}{j}\sum\limits_{k\in K_a}\alpha_k b_k \gamma_k(j) G(j-b_k)+ \\ \frac{1}{j}\sum\limits_{k\in K_a}\sum\limits_{s=1}^{s(k)}\alpha_{kr_s} b_{kr_s} \gamma_{kr_s}(j) G(j-b_{kr_s})+ \\ \frac{1}{\min(C,j)}\sum\limits_{k\in K_e}\alpha_k b_k \gamma_k(j) G(j-b_k)+ \\ \frac{1}{\min(C,j)}\sum\limits_{k\in K_e}\sum\limits_{s=1}^{s(k)}\alpha_{kr_s} b_{kr_s} \gamma_{kr_s}(j) G(j-b_{kr_s}) \\ \qquad \text{for } j=1,...,T \\ 0 \qquad \text{otherwise} \end{cases} \tag{36}$$

where: $\alpha_{kr}=\lambda_k \mu^{-1}_{kr}$ and

$$\gamma_k(j)=\begin{cases} 1 \text{ for } 1\le j\le C \text{ and } b_{kr_s}>0 \\ 1 \text{ for } 1\le j\le T \text{ and } b_{kr_s}=0 \\ 0 \text{ otherwise} \end{cases}$$

$$\gamma_{kr_s}(j)=\begin{cases} 1 \text{ for } C-b_{kr_{s-1}}+b_{kr_s}<j\le C \text{ and } s\ne s(k) \\ 1 \text{ for } C-b_{kr_{s-1}}+b_{kr_s}<j\le T \text{ and } s=s(k) \\ 0 \text{ otherwise} \end{cases}$$

If the link accommodates only elastic service-classes, then (36) is written as [12]:

$$G(j)=\begin{cases} 1 \qquad \text{for } j=0 \\ \frac{1}{\min(C,j)}\sum\limits_{k\in K_e}\alpha_k b_k \gamma_k(j) G(j-b_k)+ \\ \frac{1}{\min(C,j)}\sum\limits_{k\in K_e}\sum\limits_{s=1}^{s(k)}\alpha_{kr_s} b_{kr_s} \gamma_{kr_s}(j) G(j-b_{kr_s}) \\ \qquad \text{for } j=1,...,T \\ 0 \qquad \text{otherwise} \end{cases} \tag{37}$$

The blocking probability, $B_{kr_{s(k)}}$, of a retry service-class k call with its last bandwidth requirement, is given by:

$$B_{kr_{s(k)}}=\sum_{j=T-b_{kr_{s(k)}}+1}^{T} G^{-1}G(j) \tag{38}$$

IV. Application example - Evaluation

For evaluation, we present an application example and compare the analytical CBP with those obtained by simulation [15]. Since the reliability ranges of the measurements (95% confidence interval) are very small, we present only mean values (from 7 runs).

A link of capacity $C = 80$ b.u. accommodates three service-classes whose calls follow a Poisson process. Calls of the 1^{st} and 2^{nd} service-class are adaptive and do not retry, while calls of the 3^{rd} service-class are elastic and may retry. Their bandwidth requirements are b_1=1 b.u., b_2=2 b.u. and b_3=6 b.u., respectively. The reduced bandwidth of the 3^{rd} service-class calls, for two retrials (at most), are: $b_{3_{r1}}=5$ b.u. and $b_{3_{r2}}$=4 b.u. The call holding time is exponentially distributed with mean value $\mu_1^{-1}=\mu_2^{-1}=\mu_3^{-1}=1$. The initial values of the offered traffic-load are: $\alpha_1 = 20$ erl, $\alpha_2=6$ erl and $\alpha_3=2$ erl. For the retrials of the 3^{rd} service-class note that: $\alpha_3 b_3=\alpha_{3_{r1}} b_{3_{r1}}=\alpha_{3_{r2}} b_{3_{r2}}$. In the x-axis of all figures, we let $\alpha_3=2$ erl, while α_1, α_2 increase in steps of 1.0 and 0.5 erl, respectively. The last values are: $\alpha_1=26$ erl, $\alpha_2=9$ erl. We consider three values of T: a) $T=C=80$ b.u., where no bandwidth compression takes place and the EA-MRM gives the same CBP results with the MRM, b) $T=82$ b.u. where $r_{\min}=C/T=80/82$ and c) $T=84$ b.u. where $r_{\min}=C/T=80/84$. In Figs. 1, 2 and 3, we present the analytical and simulation CBP results of the 1st, 2nd and 3rd service-class (CBP of calls with $b_{3_{r2}}$), respectively, for all values of T. All figures show that our analytical models are: i) of absolutely satisfactory accuracy (compared to simulation) and ii) consistent, since the increase of T results in a CBP decrease, due to bandwidth compression.

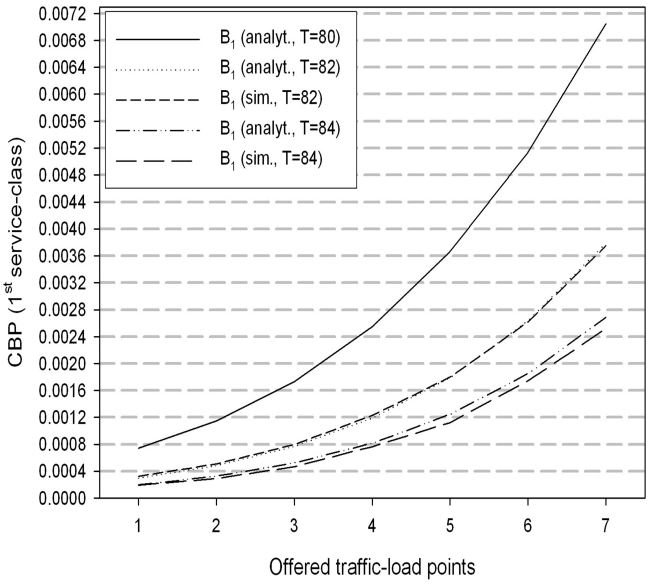


Figure 1. CBP (1^{st} service-class, adaptive).

V. Conclusion

We propose multirate loss models for a link with elastic and adaptive traffic. When Poisson arriving calls are blocked, with their initial bandwidth, have the ability to retry to be connected in the system one (EA-SRM) or more times (EA-MRM). If a retry call is blocked with its last bandwidth, it can still be accepted in the system by compressing its last bandwidth together with the bandwidth of all in-service calls. We propose approximate but recursive formulas for the efficient CBP calculation. Simulation results verify the analytical results and show that our models are accurate and consistent.

Acknowledgment

Work supported by the Research Program Caratheodory of the Research Committee of the University of Patras, Greece.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

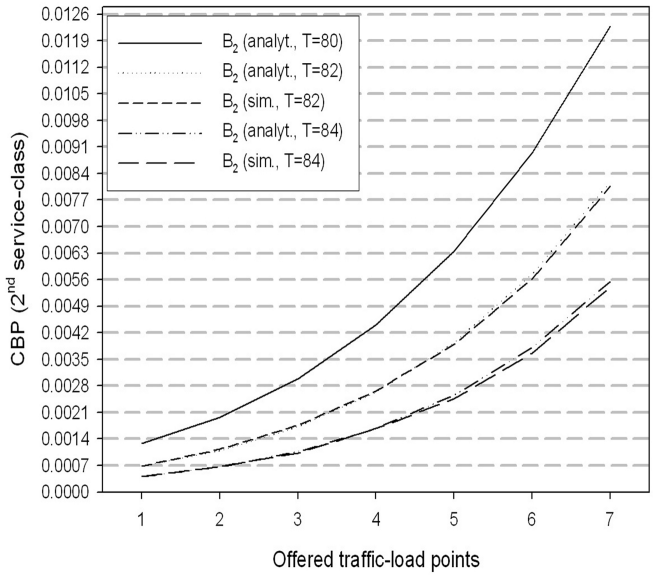


Figure 2. CBP (2^{nd} service-class, adaptive).

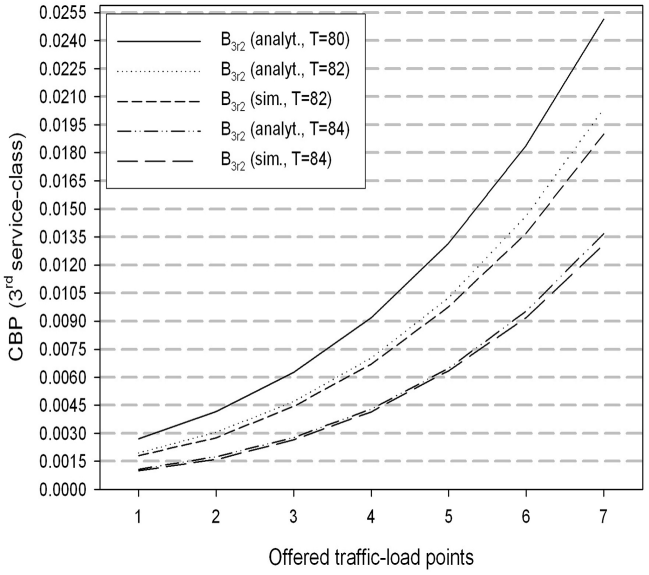


Figure 3. CBP of retry calls with b_{3r_2} (3^{rd} service-class, elastic).

REFERENCES

[1] J. S. Kaufman, "Blocking in a shared resource environment", *IEEE Trans. Commun*. vol. 29, no. 10, pp. 1474-1481, October 1981.

[2] J. W. 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.

[3] I. Moscholios, M. Logothetis, and M. Koukias, "An ON-OFF Multi-Rate Loss Model of Finite Sources", *IEICE Trans. Commun.*, Vol. E90-B, No. 7, July 2007, pp.1608-1619.

[4] Q. Huang, King-Tim Ko and V. Iversen, "Approximation of loss calculation for hierarchical networks with multiservice overflows", *IEEE Trans. Commun.*, Vol. 56, Issue 3, March 2008, pp. 466-473.

[5] I. Moscholios and M. Logothetis, "The Erlang multirate loss model with Batched Poisson arrival processes under the bandwidth reservation policy", *Computer Communications*, Vol. 33, Supplement 1, November 2010, pp. S167-S179.

[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*, Berlin, September 2003, pp. 81-90.

[7] M. Glabowski, M. Stasiak, A. Wisniewski, and P. Zwierzykowski, "Blocking Probability Calculation for Cellular Systems with WCDMA Radio Interface Servicing PCT1 and PCT2 Multirate Traffic", *IEICE Trans. Commun.*, vol.E92-B, April 2009, pp.1156-1165.

[8] J. Vardakas, V. Vassilakis and M. Logothetis, "Blocking Analysis in Hybrid TDM-WDM Passive Optical Networks", *Proc. 5th Int. Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks*, Karlskrona, Sweden, February 2008.

[9] K. Kuppuswamy, D. Lee, "An analytic approach to efficiently computing call blocking probabilities for multiclass WDM networks", *IEEE/ACM Trans. Netw.*, Vol. 17, Issue 2, April 2009, pp. 658-670.

[10] J. Vardakas, I. Moscholios, M. Logothetis and V. Stylianakis, "An Analytical Approach for Dynamic Wavelength Allocation in WDM-TDMA PONs Servicing ON-OFF Traffic", *IEEE/OSA Journal of Optical Commun. Networking*, Vol. 3, No. 4, April 2011, pp. 347-358.

[11] J. S. Kaufman, "Blocking with retrials in a completely shared resource environment", *Performance Evaluation*, Vol. 15, Issue 2, June 1992, pp. 99-113.

[12] I. Moscholios, V. Vassilakis, J. Vardakas and M. Logothetis, "Retry loss models supporting elastic traffic", *Advances in Electronics and Telecommun.*, Poznan Univ. of Technology, Poland, Vol. 2, No. 3, September 2011, pp. 8-13.

[13] S. Racz, B. Gero and G. Fodor, "Flow Level Performance Analysis of a Multi-service System Supporting Elastic and Adaptive Services", *Performance Evaluation*, Vol.49, Issues 1-4, September 2002, pp. 451-469.

[14] V. Vassilakis, I. Moscholios and M. Logothetis, "Call-level Performance Modeling of Elastic and Adaptive Service-classes", *IEEE ICC*, Glasgow, U.K., 24-28 June 2007.

[15] Simscript II.5, http://www.simscript.com.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Blocking Probabilities in Multicast WDM Optical Networks With First-Fit Wavelength Assignment

Anwar Alyatama
Computer Engineering Department
Kuwait University
P. O. Box 5969, Safat 13060, Kuwait
Email: a.yatama@ku.edu.kw

***Abstract*—We present an approximate analytical method to evaluate the blocking probabilities in multicast Wavelength Division Multiplexing (WDM) networks without wavelength converters. Our approach is based on iteratively solving the multicast call blocking probabilities for fixed routing with First-Fit wavelength assignment algorithm. We divide the WDM network into layers (colors) and we use the moment matching method to characterize the overflow traffic from one layer to another. Analyzing blocking probabilities in each layer of the network is derived from an exact approach. Results are presented which indicate the accuracy of our method.**

***Index Terms*—Blocking probability, Mutlicast Routing, WDM.**

I. Introduction

Wavelength-division multiplexing (WDM) has the potential of delivering huge bandwidth by providing many lightpaths simultaneously on one fiber. Each lightpath is independent and located at a different wavelength. A lightpath may span multiple fiber links to provide a circuit-switched interconnection between two nodes. When the network does not have conversion capabilities, the same wavelength must be available on all links. Many applications such as distribution of video require a multicast connection. A multicast connection contains a source node and a group of destination nodes. The subnetwork spanning the source node and the group of destination nodes is called a multicast tree. Using such trees, signals are transmitted to the leaf (destination) nodes in the multicast tree. Signals pass through a non-leaf destination node are dropped locally, but a copy of it is also transmitted downstream to the next node. However, finding an optimal multicast tree is not easy, and many algorithms are introduced to solve the multicast tree problem.

A challenging issue is the multicast call blocking probability in which, given a multicast call (request) traffic rate that need to be established on the network, and given a constraint on the number of wavelengths, calculate the probability that no common wavelength is available on the predetermined multicast tree. The problem of evaluating the multicast call blocking probabilities has been studied in several studies [1] [2]. They differ in their underlying assumptions and have varying computational complexities and level of accuracy. The purpose of this paper is to derive an iterative model to calculate the call blocking probabilities for fixed routing in Multicast WDM networks. Our approach uses the wavelength independence assumption. We analyze a given wavelength-routing network by dividing it into layers (colors). The analysis of each layer is derived from an exact approach. The overflow traffic from one layer to another is characterized by the moment matching method (Section III). An equivalent path method is used to calculate the overflow moments. These moments are used to calculate the equivalent Poisson overflow load used in the calculation of the path blocking probabilities. The analytical model presented in this paper is based on the work in [3] for the unicast WDM. The model is applicable to arbitrary network topologies with static routing and First-Fit wavelength assignment.

The rest of the paper is organized as follows: next section presents the network model. In section III, we introduce our proposed solution. Section IV presents some numerical results. Lastly, we present our conclusion.

II. Network Model

A call is considered the basic unit of WDM traffic. A multicast call originating from a source node s to the set of destination nodes $D_t = \{d_1, d_2, \ldots, d_i\}$ is denoted as (s, D_t). A unicast call has a single destination $|D_t| = 1$. A predetermined light-tree (i.e. an all-optical multicast tree) $T(s, D_t)$ exists for each multicast request (s, D_t) at the design stage [4]. The call arrival process entering the network is assumed to be Poisson with rate λ_{s,D_t} calls/unit time. The call termination process is exponentially distributed with a mean $\mu = 1$. The arrival and termination rates are assumed to be equal. The call requires one wavelength (channel) to be available from each link along the predetermined fixed tree $T(s, D_t)$ from the source s to each destination $d \in D_t$. Since no conversion capability is assumed the same wavelength must be used in all links belonging to the tree; otherwise the call request is blocked. The nodes in the network are classified into two categories: split incapable or split capable. Multicast split incapable nodes are nodes which cannot split the incoming lightpath to more than one output link. However, the implementation of a split capable node may be expensive due to the large amount of amplification and fabrication [5]. All nodes of the network have Drop-and-Continue capability [6].

Let the order of wavelengths be numbered $w = 1, 2, 3, \ldots, W$. Upon the arrival of a call, the source s will

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

offer the call to the first wavelength (layer) $w = 1$ on the predetermined fixed tree. The call is accepted if the wavelength w is available on all links belonging to the predetermined fixed path. Otherwise, the call is offered to the next wavelength. Thus, the traffic which cannot be carried by a wavelength w is offered to the next wavelength $w+1$ and so on until the call is either accepted or blocked. Multicast call (s, D_t) blocking probability for a given wavelength w is denoted by P^w_{s,D_t}. The link i, j capacity is denoted as $C_{i,j}$ and the path $r(s, d)$ capacity is denoted as $C_{s,d}$, where $C_{s,d} = \min_{(i,j)\in r(s,d)} C_{i,j}$. Similarly, C_{s,D_t} denotes the capacity of tree $T(s, D_t)$.

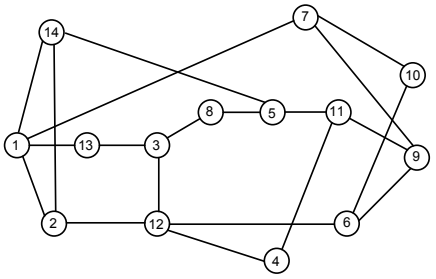


Fig. 1. The 14-Nodes NSFNET network topology

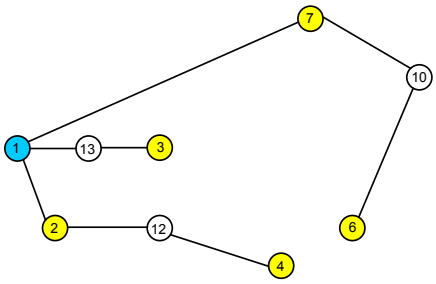


Fig. 2. A multicast tree $T(1, D_1)$, $D_1 = \{2, 3, 4, 6, 7\}$. Transient nodes are Drop-and-Continue capable but not splitting capable. The tree $T(1, D_1)$ contains the route $r(1, 3), r(1, 4), r(1, 6)$.

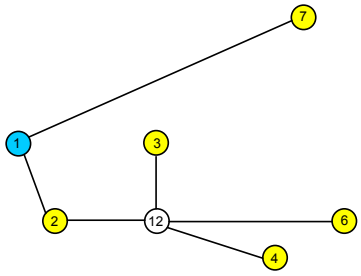


Fig. 3. A multicast tree $T(1, D)$, $D = \{2, 3, 4, 6, 7\}$. The tree $T(1, D)$ contains the route $r(1, 7), r(1, 12)$ and subtree $T(12, \{3, 4, 6\})$. Node 12 is a transient and split capable node.

Fig. 1 shows the NSFNET network topology. Fig. 2 and Fig. 3 show a multicast call $T(s, D)$, $D = \{2, 3, 4, 6, 7\}$ with two trees. All nodes in Fig. 2 are split incapable nodes, whereas node 12 is a split capable node in Fig. 3. The first tree can be descried by the set of paths $T(s, D) = r(s, d_i)$ and $r(s, d_i) \nsubseteq r(s, d_j)$. The second tree is described by the set of paths $r(s, d_i)$ and $r(d_s, d_j)$, where node d_s is a split capable node.

III. Proposed Solution

In this section, we present a basic description of the wavelength decomposition method for unicast calls introduced in [3]. Wavelength decomposition method analyzes the network by splitting it into layers and uses a moment matching method to calculate an equivalent Poisson overflow traffic to each layer. The analysis of blocking probabilities in each layer is derived from an exact approach. In this work, we extend the single layer unicast blocking probability calculations to solve the multicast call blocking probability.

A. The Single Layer Blocking Probability

First, we review the basic concept of calculating the single layer blocking probability for unicast calls in [3]. Consider the $k-1$ hop route shown in Fig. 4, denoted as $r(1, k)$. Let the state of path $r(1, k)$ in a single wavelength (layer) w at time t be described by the $k(k-1)/2$ dimensional process

$$X^w_{r(1,k)}(t) = (n^w_{1,2}(t), n^w_{1,3}(t), \ldots, n^w_{k-1,k}(t)) \tag{1}$$

The state of the $k-1$ hop path $r(1, k)$ for wavelength w is thus denoted by the number of calls $n^w_{i,j} \in \{0, 1\}$ in progress for each segment $r(i, j)$, $1 \le i < k$, $1 < j \le k$, $i < j$, where

$$\begin{gathered} n^w_{i,j} + n^w_{l,m} \le 1 \quad \forall\, r(i,j) \cap r(l,m) \neq \emptyset \quad \text{and} \\ 1 \le l < k, 1 < m \le k, l < m \end{gathered} \tag{2}$$

Process $X^w_{r(1,k)}(t)$ is a time-reversible Markov process and the stationary vector π is given by [7]

$$\begin{aligned} \pi(n^w_{1,2}, n^w_{1,3}, \ldots, n^w_{k-1,k}) = \frac{1}{G^w_{r(1,k)}} \Bigg[& \frac{(a^w_{1,2})^{n^w_{1,2}}}{n^w_{1,2}!} \cdot \\ & \frac{(a^w_{1,3})^{n^w_{1,3}}}{n^w_{1,3}!} \cdots \frac{(a^w_{k-1,k})^{n^w_{k-1,k}}}{n^w_{k-1,k}!} \Bigg] \end{aligned} \tag{3}$$

where $a^w_{i,j}$ is the background traffic in each segment i, j for a given wavelength w. $G^w_{r(1,k)}$ is the normalization constant for wavelength w on the path $r(1, k)$ and is given by

$$G^w_{r(1,k)} = \sum_{\substack{n^w_{i,j}+n^w_{l,m}\le 1 \\ \forall\, r(i,j)\cap r(l,m)\neq\emptyset \\ r(i,j)\subseteq r(s,d),\, r(l,m)\subseteq r(1,k)}} \prod_{r(i,j)\subseteq r(1,k)} \frac{(a^w_{i,j})^{n^w_{i,j}}}{n^w_{i,j}!} \tag{4}$$

by using the reduced blocking path model [8], the background traffic $a^w_{i,j}$ for segment $r(i, j) \subseteq T(s, D_t)$ is calculated as [3],

$$a^w_{i,j} = \sum_{\substack{r(i,j)\subseteq T(s,D_t) \\ r(i,j)\leftarrow A^w_{s,D} \text{ then } r(m,l)\nleftarrow A^w_{s,D_t} \\ \forall\ r(l,m)\subseteq r(1,k)}} \frac{A^w_{s,D_t} \cdot (1 - P^w_{s,D_t})}{1 - P^w_{i,j}} \tag{5}$$

where A^w_{s,D_t} is the source s to group D_t offered load at wavelength w, $A^1_{s,D_t} = \lambda_{s,D_t}$. The normalization constant $G^w_{r(1,k)}$ can be calculated recursively [3] as

$$G^w_{r(1,k)} = G^w_{r(1,k-1)} + \sum_{i=1}^{k-1} G^w_{r(1,k-i)} a^w_{k-i,k} \tag{6}$$

where $G^w_{r(1,1)} = 1$. Thus, path $r(1,k)$ blocking probability $P^w_{r(1,k)}$ *(or $P^w_{1,k}$ for short)* in a single layer $w \leq C_{1,k}$ is calculated as

$$P^w_{r(1,k)} = 1 - \pi(0,0,\ldots,0) = 1 - \frac{1}{G^w_{r(1,k)}} \tag{7}$$

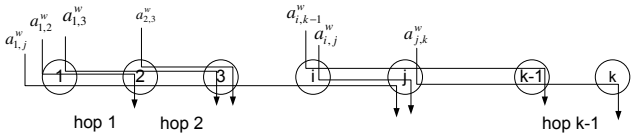


Fig. 4. A $k-1$ hop path $r(1,k)$. The state of the path in wavelength (layer) w at time t is described by the $k(k-1)/2$ dimensional process $X^w_{r(1,k)}(t) = (n^w_{1,2}(t), n^w_{1,3}(t), \ldots, n^w_{k-1,k}(t))$. Where $n^w_{i,j}(t)$ is the number of calls using the segment $r(i,j)$, that are currently active in wavelength w at time t i.e., $n^w_{i,j}(t) \in \{0,1\}$. The background offered traffic in each segment i,j for a given wavelength w is denoted as $a^w_{i,j}$.

Now, to extend the unicast wavelength decomposition method to multicast WDM network let us consider two disjoint routes *(paths)* $r(s_1,d_1)$ and $r(s_2,d_2)$ as shown in figure 5. Common traffic passing through both routes are denoted as $a^w_{(i,j),(x,y)}$ where $i,j \in r(s_1,d_1)$ and $x,y \in r(s_2,d_2)$. We can calculate the single layer joint normalization constant for the two routes $G^w_{r(s_1,d_1)} \cap G^w_{r(s_2,d_2)}$ as follow.

$$\begin{aligned} G^w_{r(s_1,d_1)} \cap G^w_{r(s_2,d_2)} &= g^w_{r(s_1,d_1)} \cdot g^w_{r(s_2,d_2)} \\ &+ \sum_{\substack{\forall\, i,j \in r(s_1,d_1) \\ \forall\, x,y \in r(s_2,d_2)}} g^w_{r(s_1,i)} \cdot g^w_{r(j,d_1)} \cdot a^w_{(i,j),(x,y)} \cdot g^w_{r(s_2,x)} \cdot g^w_{r(y,d_2)} \end{aligned} \tag{8}$$

where, $g^w_{r(i,j)}$ represents the normalization constant for segment $r(i,j)$ excluding all common traffic between route $r(s_1,d_1)$ and route $r(s_2,d_2)$.

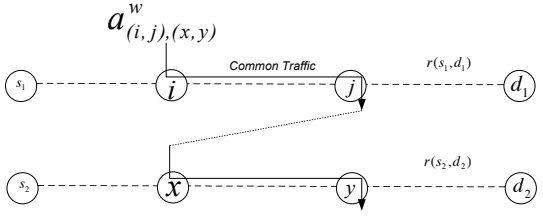


Fig. 5. Two disjoint routes *(paths)* $r(s_1,d_1)$ and $r(s_2,d_2)$

Consider a tree $T(s,D)$ (e.g., Fig. 2) for a given wavelength w. The event that a wavelength w is available to the multicast call $T(s,D)$ is conditioned on the availability of wavelength w on all links belonging to the tree. Therefore, a multicast call (s,D_t) blocking probability for a given wavelength w is given by

$$P^w_{s,D_t} = 1 - \frac{1}{\mathbf{G}^w_{s,D_t}} \tag{9}$$

where, $\mathbf{G}^w_{s,D_t}$ is the multicast call $T(s,D_t)$ normalization constant. If all nodes are split incapable, then the tree is a group of disjoint paths $r(s,d_i)$. Hence,

$$\mathbf{G}^w_{s,D_t} = \bigcap_{\forall i} G^w_{r(s,d_i)} + A^w_{s,D_t} \tag{10}$$

For example, the normalization constant $\mathbf{G}^w_{1,D_1}$ for the tree in Fig. 2 is

$$\mathbf{G}^w_{1,\{2,3,4,6,7\}} = G^w_{r(1,3)} \cap G^w_{r(1,4)} \cap G^w_{r(1,6)} + A^w_{1,\{2,3,4,6,7\}}$$

Since $r(1,2) \subseteq r(1,4)$ and $r(1,7) \subseteq r(1,6)$.

For networks with split capable nodes, trees will be a mix of disjoint paths with no split capable node and subtrees. The split capable nodes d_s will be the root of these subtrees and the leaves will be a subset $D_{t'} \subseteq D_t$. Therefore, the normalization constant will be the product of disjoint paths $G^w_{r(s,d_s)}$ and subtrees $\mathbf{G}^w_{d_s,D_{t'}}$ calculated from Eq. 12. Now, consider a branch with a split capable node d_s to a subset $D_{t'}$ as shown in Fig. 6. The normalization constant, $\mathbf{G}^w_{d_{s-1},D_{t'}}$ can be calculated form $\mathbf{G}^w_{d_s,D_{t'}}$ as

$$\begin{aligned} \mathbf{G}^w_{d_{s-1},D_{t'}} &= G^w_{r(d_{s-1},d_s)} \cdot \mathbf{G}^w_{d_s,D_{t'}} + \sum_{\forall k \in D_{t'}} \\ &\sum_{\forall j \in r(d_{s+1},k)} a^w_{d_{s-1},j}\, \mathbf{G}^w_{d_s,D_{t'}-\{k\}} + A^w_{d_{s-1},D_{t'}} \end{aligned}$$

Generally, the normalization constant, $\mathbf{G}^w_{s,D_{t'}}$ is

$$\begin{aligned} \mathbf{G}^w_{s,D_{t'}} &= G^w_{r(d_i,d_{i+1})} \cdot \mathbf{G}^w_{d_{i+1},D_{t'}} \\ &+ \sum_{\forall\, i \in r(s,d_s-1)} \sum_{\forall k \in D_{t'}} \sum_{\forall j \in r(d_{s+1},k)} G^w_{r(s,i)} \\ &\cdot\ a^w_{i,j}\ \ \mathbf{G}^w_{d_s,D_{t'}-\{k\}} + A^w_{s,D_{t'}} \end{aligned} \tag{11}$$

Finally, $\mathbf{G}^w_{s,D_t}$

$$\mathbf{G}^w_{s,D_t} = \prod_{\forall D_{t'}} \mathbf{G}^w_{s,D_{t'}} \cdot \prod_{\substack{\forall d_i \in D_t, d_i \notin D_{t'} \\ r(s,d_i) \nsubseteq r(s,d_j)\, \forall\, d_j \in D_t}} G^w_{r(s,d_i)} + A^w_{s,D_t} \tag{12}$$

For example, the normalization constant $\mathbf{G}^w_{1,D}$ for the tree

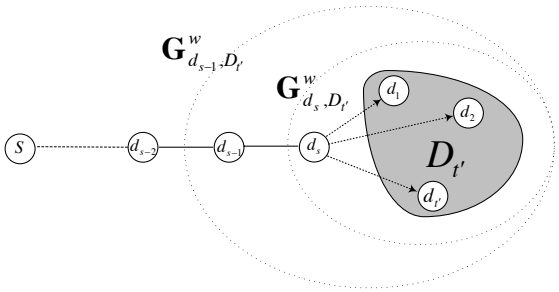


Fig. 6. Calculating the normalization constant recursively, for a multicast tree with a split capable node d_s.

in Fig. 3 is

$$\mathbf{G}^w_{1,\{2,3,4,6,7\}} = G^w_{r(1,7)} \cap \mathbf{G}^w_{1,\{3,4,6\}} + A^w_{1,D}$$

$$\begin{aligned}\mathbf{G}^w_{1,\{3,4,6\}} &= G^w_{r(1,12)} \cdot \mathbf{G}^w_{12,\{3,4,6\}} \\ &+ G^w_{r(1,1)} \cdot a^w_{1,3} \cdot \mathbf{G}^w_{12,\{4,6\}} + G^w_{r(1,1)} \cdot a^w_{1,4} \cdot \mathbf{G}^w_{12,\{3,6\}} \\ &+ G^w_{r(1,1)} \cdot a^w_{1,6} \cdot \mathbf{G}^w_{12,\{3,4\}} + G^w_{r(1,2)} \cdot a^w_{2,3} \cdot \mathbf{G}^w_{12,\{4,6\}} \\ &+ G^w_{r(1,2)} \cdot a^w_{2,4} \cdot \mathbf{G}^w_{12,\{3,6\}} + G^w_{r(1,2)} \cdot a^w_{2,6} \cdot \mathbf{G}^w_{12,\{3,4\}}\end{aligned}$$

To simplify notations we will drop the subscript t from D_t.

B. Calculating the Moments of the Overflow Traffic

Since the overflow load from wavelength $w+1$ is non Poisson, we need to calculate both the first and the second overflow traffic moments (*mean* $\bar{A}^{w+1}$ *and variance* $\bar{V}^{w+1}$) to the next layer $w+1$. For this, we construct an equivalent single-link system such that the blocking of the Poisson traffic in this system will approximate the blocking on the tree $T(s,D)$. We know that the total offered load to the tree is $\lambda_{s,D}$ and the overflow mean, up to the current wavelength w is $A^w_{s,D}P^w_{s,D}$. Hence,

$$\lambda_{s,D} \cdot E_r(\lambda_{s,D}, N^w_{s,D}) = A^w_{s,D} \cdot P^w_{s,D} \tag{13}$$

where, E_r is the generalized (not integral) Erlang-B formula [9]. The overflow mean to wavelength $w+1$ is

$$\bar{A}^{w+1}_{s,D} = \lambda_{s,D} \cdot E_r(\lambda_{s,D}, N^w_{s,D}) \tag{14}$$

The variance $\bar{V}^{w+1}_{s,D}$ is calculated using Riordan's formula [8],

$$\bar{V}^{w+1}_{s,D} = \bar{A}^{w+1}_{s,D}\left(1 - \bar{A}^{w+1}_{s,D} + \frac{\lambda_{s,D}}{N^w_{s,D} + 1 + \bar{A}^{w+1}_{s,D} - \lambda_{s,D}}\right) \tag{15}$$

The peakedness is defined as, $\bar{Z}^{w+1}_{s,D} = \frac{\bar{V}^{w+1}_{s,D}}{\bar{A}^{w+1}_{s,D}}$.

C. Calculating the Equivalent Poisson Traffic

The path blocking probabilities calculated in Eq. 5 through Eq. 12 assume Poisson traffic with $\bar{Z}^w_{s,D} = 1$. However, the overflow traffic to layer $w+1$ calculated from equations 14 and 15 is in general non Poisson $\bar{Z}^{w+1}_{s,D} \neq 1$. We again use an equivalent single-link system with $\bar{N}^{w+1}_{s,D} \leq 1$ wavelengths to find an equivalent Poisson traffic with mean $A^{w+1}_{s,D}$ and $Z^{w+1}_{s,D} = 1$ that matches the overflow traffic with mean $\bar{A}^{w+1}_{s,D}$ and variance $\bar{V}^{w+1}_{s,D}$.

Fredricks and Hayward's equivalence method was used with the original wavelength decomposition method described in [3]. It attempts to describe a non-Poisson traffic $Z \neq 1$ by an equivalent Poisson traffic $Z = 1$ [9]. Mainly, the blocking probability of the actual system with $\bar{N}^w_{s,d}$ channels offered non-Poisson traffic with rate $\bar{A}^w_{s,d}$ and peakedness $\bar{Z}^w_{s,d} \neq 1$ has the same blocking probability with $\bar{N}^w_{s,d}/\bar{Z}^w_{s,d}$ channels, offered $\bar{A}^w_{s,d}/\bar{Z}^w_{s,d}$ traffic, and peakedness $Z^w_{s,d} = 1$ (Poisson).

In this work, we combine Fredricks and Hayward's equivalence method with Berkeley's Equivalent Random Traffic (ERT) approximation. Combining moment matching functions seems to be more suitable to calculate the equivalent Poisson traffic [10]. The idea of the Equivalent Random Traffic (ERT) method is to think that the traffic with mean $\bar{A}^w_{s,d}$ and variance $\bar{V}^w_{s,d} \neq \bar{A}^w_{s,d}$ is obtained as overflow traffic from a fictitious system with $\hat{\mathbf{N}}^w_{s,d}$ channels offered a Poisson traffic with mean $\hat{\mathbf{A}}^w_{s,d}$. Hence, the blocking probability of the non-Poison secondary system is the same as the blocking probability of the equivalent system with $\hat{\mathbf{N}}^w_{s,d}$ channels, mean $\hat{\mathbf{A}}^w_{s,d}$ and peakedness $Z^w_{s,d} = 1$. Berkeley's ERT approximation is considered as a single parameter ERT method where, we fix $\hat{\mathbf{A}}^w_{s,d} = \lambda_{s,d}$ [8] [11]. Hence, $\hat{\mathbf{N}}^w_{s,d}$ is as follows

$$\hat{\mathbf{N}}^w_{s,d} = \frac{\lambda_{s,d}(\bar{A}^w_{s,d} + \bar{Z}^w_{s,d})}{\bar{A}^w_{s,d} + \bar{Z}^w_{s,d} - 1} - \bar{A}^w_{s,d} - 1 \tag{16}$$

D. Calculating the Overall Path Blocking Probability

The overall path blocking probability is calculated as

$$P_{s,D} = \frac{A^{C_{s,D}}_{s,D} \cdot P^{C_{s,D}}_{s,D}}{\lambda_{s,D}} \tag{17}$$

IV. NUMERICAL RESULTS

First, we present a five node network shown in Fig. 7 as an illustrative example for First-Fit WA. There are two multicast calls, $T(1,D_1)$, $D_1 = \{3,5\}$ and $T(2,D_2)$, $D_2 = \{3,4\}$ with arrival rate of $\lambda_{1,D_1} = \lambda_{1,D_2} = 0.5$ (calls/unit time). The unicast call arrival rates are $\lambda_{1,2} = \lambda_{1,3} = \lambda_{1,4} = \lambda_{1,5} = \lambda_{2,3} = \lambda_{2,4} = 0.5$ (calls/unit time). Link capacities are 4 channels.

The normalization constant for the multicast tree $T(1,D_1)$ is

$$\begin{aligned}\mathbf{G}^w_{1,D_1} &= G^w_{r(1,3)} * G^w_{r(1,5)} + A^w_{1,D_1} \\ &= (1 + a^w_{1,2} + a^w_{2,3} + a^w_{1,2}a^w_{2,3} + A^w_{1,3}) * \\ &\quad (1 + A^w_{1,5}) + A^w_{1,D_1}\end{aligned}$$

where

$$a^w_{1,2} = \frac{A^w_{1,2}(1-P^w_{1,2}) + A^w_{1,4}(1-P^w_{1,4})}{1-P^w_{1,4}}$$

$$a^w_{2,3} = \frac{A^w_{2,3}(1-P^w_{2,3}) + A^w_{2,D_2}(1-P^w_{2,D_2})}{1-P^w_{2,3}}$$

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

The normalization constant for the multicast tree $T(2, D_2)$ is

$$\mathbf{G}^w_{2,D_2} = 1 + a^w_{2,3} + a^w_{2,4} + a^w_{2,3}a^w_{2,4} + A^w_{2,D_2}$$

where

$$a^w_{2,3} = \frac{A^w_{2,3}(1-P^w_{2,3})+A^w_{1,3}(1-P^w_{1,3})+A^w_{1,D_1}(1-P^w_{1,D_1})}{1-P^w_{2,3}}$$

$$a^w_{2,4} = \frac{A^w_{2,4}(1-P^w_{2,4})+A^w_{1,4}(1-P^w_{1,4})}{1-P^w_{2,4}}$$

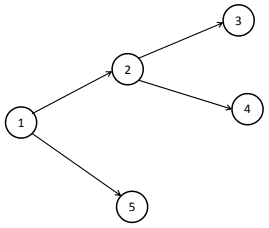


Fig. 7. A network example with five nodes

The path blocking probabilities in the first layer and the overall blocking probabilities are shown in table I.

TABLE I
THE PATH BLOCKING PROBABILITIES IN THE FIRST LAYER AND THE OVERALL BLOCKING PROBABILITIES FOR AN EXAMPLE OF A TREE WITH SPLIT INCAPABLE NODES.

layer	First		Overall	
Path	Simulation	Calculation	Simulation	Calculation
$P^1_{1,2}$	0.5519	0.5559	0.0664	0.0633
$P^1_{1,3}$	0.7557	0.7512	0.1744	0.1672
$P^1_{1,4}$	0.7557	0.7469	0.1613	0.1531
$P^1_{1,5}$	0.3887	0.3886	0.00814	0.0060
$P^1_{2,3}$	0.5517	0.5559	0.0663	0.0633
$P^1_{2,4}$	0.4977	0.5021	0.0302	0.0306
P^1_{1,D_1}	0.8376	0.8342	0.2032	0.2159
P^1_{2,D_2}	0.7557	0.7469	0.1615	0.1531

Now, let us assume node 2 is split capable node and $D_1 = \{3, 4, 5\}$ for the multicast call $T(1, D_1)$. The normalization constant for the multicast tree $T(1, \{3, 4, 5\})$ is

$$\begin{aligned}\mathbf{G}^w_{1,\{3,4,5\}} &= G^w_{r(1,5)} \cdot \mathbf{G}^w_{1,\{3,4\}} + A^w_{1,D_1}\\ \mathbf{G}^w_{1,\{3,4\}} &= G^w_{r(1,2)} \cdot \mathbf{G}^w_{2,\{3,4\}}\\ &\quad + G^w_{r(1,1)} \cdot a^w_{1,3} \cdot \mathbf{G}^w_{2,\{4\}}\\ &\quad + G^w_{r(1,1)} \cdot a^w_{1,4} \cdot \mathbf{G}^w_{2,\{3\}}\end{aligned}$$

The normalization constant for the multicast tree $T(2, D_2)$ is

$$\begin{aligned}\mathbf{G}^w_{2,D_2} &= 1 + a^w_{2,3} + a^w_{2,4} + a^w_{2,3}a^w_{2,4}\\ &\quad + A^w_{1,D_1} * (1 - P^w_{1,D_1})/(1 - P^w_{2,D_2}) + A^w_{2,D_2}\end{aligned}$$

where $a^w_{2,3}$ is

$$a^w_{2,3} = \frac{A^w_{2,3}(1-P^w_{2,3})+A^w_{1,3}(1-P^w_{1,3})}{1-P^w_{2,3}}$$

TABLE II
THE PATH BLOCKING PROBABILITIES IN THE FIRST LAYER AND THE OVERALL BLOCKING PROBABILITIES FOR AN EXAMPLE OF A TREE WITH SPLIT CAPABLE NODES.

layer	First		Overall	
Path	Simulation	Calculation	Simulation	Calculation
$P^1_{1,2}$	0.5362	0.5421	0.0639	0.0581
$P^1_{1,3}$	0.7469	0.7428	0.1686	0.1583
$P^1_{1,4}$	0.7459	0.7428	0.1685	0.1583
$P^1_{1,5}$	0.3684	0.3706	0.0083	0.0046
$P^1_{2,3}$	0.5361	0.5421	0.0638	0.0581
$P^1_{2,4}$	0.5370	0.5421	0.0641	0.0581
P^1_{1,D_1}	0.8862	0.8881	0.2478	0.2660
P^1_{2,D_2}	0.7467	0.7428	0.1688	0.1583

The path blocking probabilities in the first layer and the overall blocking probabilities are shown in table II.

Our next test vehicle is the 14 nodes NSFNET as shown in Fig. 1. The network traffic for unicast calls is similar to a realistic network with realistic traffic, and the network has been dimensioned using shortest path [12]. The multicast traffic for split incapable nodes is given in Table III. For split capable network, we arbitrary chose nodes 12 as split capable node. Fig. 3 shows the new tree for the multicast call $D_1 = \{2, 3, 4, 6, 7\}$.

Simulation results are run 10 times and each run starts with a different random seed where, each seed simulation runs for 10,000 holding times. The overall average result is obtained with 95% confidence. For the analytical techniques, the iterative algorithm terminates when all blocking probability values have converged within 10^{-5}.

TABLE III
MULTICAST TRAFFIC WITH LOAD FACTOR=12 FOR SPLIT INCAPABLE NODES.

Source	D	$\lambda_{s,D}$	Routes	Links
1	2,3,4,6,7	0.1	$r(1,3), r(1,4), r(1,6)$	8
1	7,10,12	4.0	$r(1,10), r(1,12)$	4
2	3,5,14	2.5	$r(2,3), r(2,5)$	4
4	1,9,11	5.0	$r(4,1), r(4,9)$	5
5	2,6,9,10	1.5	$r(5,2), r(5,10)$	6
6	2,5,10,11	7.0	$r(6,12), r(6,5), r(6,10)$	6
7	1,4,9,14	0.5	$r(7,4), r(7,14)$	5
8	1,5,13,14	2.0	$r(8,13), r(8,14)$	4
9	8,6	2.0	$r(9,6), r(9,8)$	4
10	1,4,12	4.0	$r(10,1), r(10,4)$	5
12	1,2,11	5.0	$r(12,1), r(12,11)$	4
12	3,8,10	4.0	$r(12,8), r(12,10)$	4
13	3,4,7,10,11	0.1	$r(13,10), r(13,11)$	7
14	2,3,5,8,12	0.25	$r(14,3), r(14,12)$	5

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

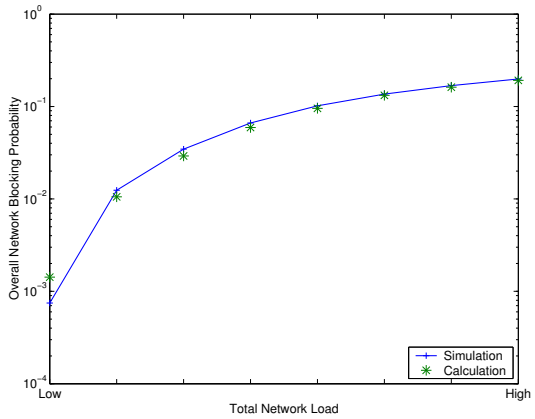


Fig. 8. Overall network end-to-end blocking probability for the NSFNET mesh network with no split capability.

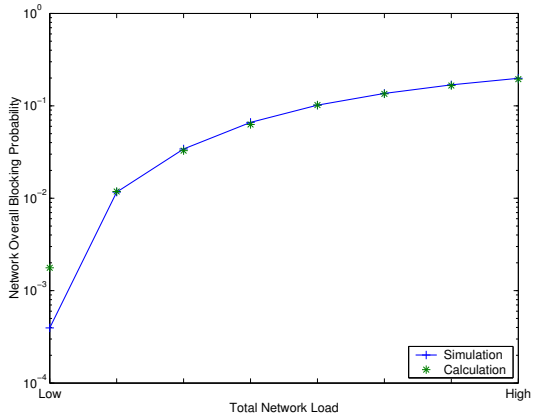


Fig. 9. Overall network end-to-end blocking probability for the NSFNET mesh network with a split capable node 12.

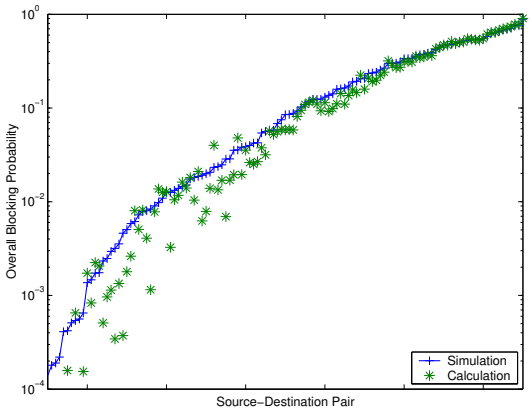


Fig. 10. The end-to-end blocking probabilities for various unicast calls in the NSFNET mesh network with no split capability.

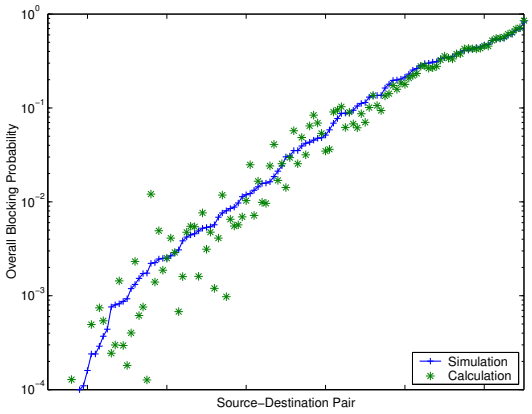


Fig. 11. The end-to-end blocking probabilities for various unicast calls in the NSFNET mesh network with a split capable node 12.

In Figs. 8 and 9, we plot the overall network blocking probability against the total network load for the split incapable and the split capable network respectively. In both figures, the first curve is obtained from simulation, and the second curve is plotted from our new approach. Figs. 10 and 11 show the end-to-end blocking probabilities for various unicast calls for both cases. The source/destination pairs are numbered in ascending order of their blocking probability values obtained from simulations. There are 170 source/destination pairs that have none zero load. Source/destination pair that yields an end-to-end blocking probability of at least 10^{-4} in the simulation is shown in the figure. Similarly, Fig. 12 and Fig. 13 show the end-to-end blocking probabilities for various multicast calls in the network. Again, trees are numbered in ascending order of their blocking probability values obtained from simulations. We can notice that the outputs of our calculations are very close to simulation results.

V. The Conclusion

We have presented a new analytical approach to evaluate more accurately the call blocking probabilities of a multicast Wavelength Division Multiplexing (WDM) network. Our approach assumes fixed routing, First-Fit wavelength assignment with split incapable or capable nodes. The new approach views the WDM network as a set of different layers (colors) where, blocked traffic in one layer is overflowed to another layer. Simulation results show the accuracy of our approach. Applying our new method to random wavelength assignment

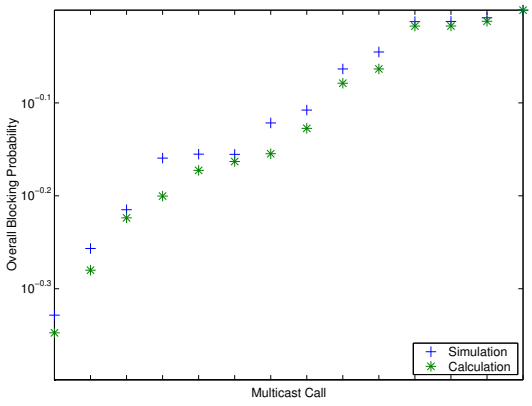


Fig. 12. The end-to-end blocking probabilities for various multicast calls in the NSFNET mesh network.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

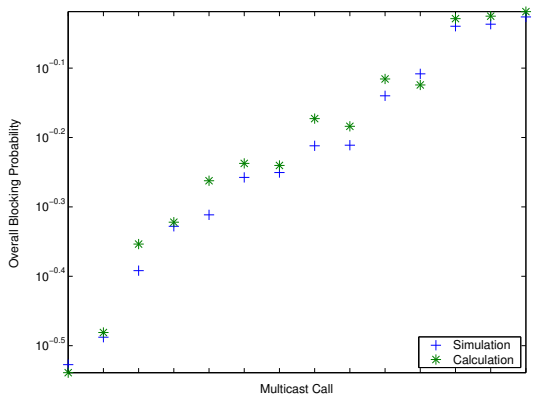


Fig. 13. The end-to-end blocking probabilities for various multicast calls in the NSFNET mesh network with a split capable node 12.

in multicast networks will be considered in future research.

REFERENCES

[1] A. Kamal and A. Alyatama, "Blocking probabilities in circuit-switched wavelength division multiplexing networks under multicast service," *Performance Evaluation*, vol. 47, no. 1, pp. 43-71, 2002.

[2] S. Ramesh, G. Rouskas and H. Perros, "Computing blocking probabilities in multiclass wavelength routing network," *IEEE Journal on Selected Areas in Communications*, vol. 20, Jan., 2002.

[3] A. Alyatama, "Wavelength decomposition approach for computing blocking probabilities in WDM optical networks without wavelength conversions," *Computer Networks*, vol. 49, pp. 727-742, Dec., 2005.

[4] J. He, S. Chan, and D. Tsang, "Multicasting in WDM Networks," *EEE Communications Surveys & Tutorials*, vol. 4, no. 1, 2002.

[5] J. Siregar, Y. Zhang and H. Takagi, "Optimal multicast routing using genetic algorithm for WDM optical networks," *IEEE Transactions on Communications*, vol. E88B, no. 1, pp. 219-226, Jan., 2005.

[6] R. Ramaswami and K. Sivarajan, *Optical Nerworks: A practical Prospective*, Second Edition, Morgan Kaufmann.

[7] C. Lea and A. Alyatama, "Bandwidth quantization and states reduction in the broadband ISDN," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, June, 1995.

[8] A. Girard, *Routing and Dimensioning in Circuit-Switched Networks*, Addison-Wesley, 1990.

[9] ITU-D Study Group 2, *Teletraffic Engineering Handbook*, Geneva, Dec., 2002.

[10] A. Alyatama, "Computing the equivalent Poisson traffic for the Wavelength Decomposition Method", *Journal of High Speed Networks*, vol. 15, no. 4, pp. 399-414, 2006.

[11] E. Karasan and E. Ayanoglu, "Effects of wavelength routing and selection algorithms on wavelength conversion gain in WDM optical networks," *IEEE/ACM Transactions on Networking*, vol. 6, no. 2, April, 1998.

[12] R. Hulsermann, M. Jager, S. Krumke, D. Poensgen, J. Ramboui and A. Turchscherer, "Dynamic routing algorithms in transparent optical networks," *Proceedings of ONDM 2003*, pp. 293-312, Feb., 2003, Budapest, Hungary.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Modelling Limited-availability Systems with Multi-service Sources and Bandwidth Reservation

Mariusz Głąbowski
Poznań University of Technology
Chair of Communication
and Computer Networks
Email: mariusz.glabowski@put.poznan.pl

Maciej Sobieraj
Poznań University of Technology
Chair of Communication
and Computer Networks
Email: maciej.sobieraj@put.poznan.pl

Maciej Stasiak
Poznań University of Technology
Chair of Communication
and Computer Networks
Email: stasiak@et.put.poznan.pl

***Abstract*—The aim of this paper is to present a new analytical calculation method for the occupancy distribution and the blocking probability in the so-called limited-availability group with multi-service sources and reservation mechanisms. The paper considers multi-service limited-availability systems with multi-service sources, in which each single traffic source can generate calls of different traffic classes. To date, only models of multi-service systems with single-service sources, in which a single source of a given class generates always only calls of this particular class have been considered in teletraffic literature. The results of analytical modeling of the limited-availability systems with multi-service sources and reservation mechanisms are compared with simulation data, which confirm a high accuracy of the method. Any possible application of the proposed model can be considered in the context of wireless networks with multi-service sources and reservation mechanisms, as well as in the context of switching networks.**

***Keywords*-limited-availability systems; multi-service networks; multi-service sources; bandwidth reservation.**

I. INTRODUCTION

Cellular networks are one of the most rapidly growing areas of telecommunications and one of the most popular systems of mobile communication. They can be used for voice transmission, but are also very efficient for sending data streams from different applications [1] [2] [3]. Data transmission is a type of service that originally was not handled by cellular networks. Over time, data transmission has become more and more popular in expanding mobile networks. Data transmission services offered by operators include video conferencing services, streaming audio services, electronic mail and large file transmission [4] [5] [6] [7].

The increase in traffic intensity of traffic generated by data transmission services is accompanied by an increase in the requirements with respect to the size of resources offered by networks. This also causes a growing necessity of working out mechanisms that introduce differentiation in Quality of Service (QoS) for particular data classes. The introduction of QoS differentiation mechanisms was conducive in turn to a development of new analytical models for dimensioning of multi-service mobile networks. The initial models of multi-service cellular networks (cell groups models) assumed that a single traffic source of a given class could generate only one, strictly defined, type of the traffic stream (traffic sources class unequivocally determined the nature of a traffic stream). Both cell groups without QoS mechanisms introduced [8], and cell groups with QoS differentiation mechanisms, were considered, including resource reservation mechanisms [9] [10] [11].

With multi-service terminals becoming more and more universal in modern cellular networks, it has become necessary to develop new traffic models. In [12], a model of the multi-service network was presented for the first time, which assumed that a given and defined set of services was related to a single traffic source. The considered system was described as a multi-service system with multi-service sources. The considerations presented in [12] were limited, however, to a model of a single full-availability cell (single resource) without any Quality of Service differentiation mechanisms introduced.

This paper proposes a model of a limited-availability group of resources, in which — to facilitate resource usage — a reservation mechanism has been implemented. The model can be used, similarly as in [12], for modeling cell groups in multi-service networks. In the proposed model, it is taken into consideration that a single terminal can generate various traffic streams corresponding to particular services implemented in the terminal. Additionally, it is assumed that the services cannot be used simultaneously by the terminal. This means that when the terminal is involved in the generation of traffic stream related to one service, it cannot at the same time generate traffic streams related to other services.

The remaining part of the paper is organized as follows. In Section II, the analytical model of the limited-availability systems with multi-service sources and resource reservation mechanism is proposed. In Section III, the results of the blocking probability obtained for three limited-availability systems with multi-service sources and reservation mechanisms are compared with the simulation data. Section IV concludes the paper and presents further work.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

II. Analytical Model of Limited-Availability Group with Multi-service Sources and Reservation Mechanisms

Let us consider a model of the limited-availability group (LAG) with multi-service sources and the capacity V_L, presented in Figure 1. The limited-availability group consists of υ identical separated resources (e.g., transmission links) with the capacity equal to f BBUs (basic bandwidth units, i.e., allocation units). The total capacity of the system V_L is equal to $V_L = \upsilon f$ BBUs.

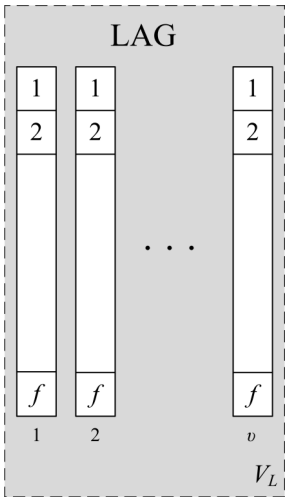


Figure 1. Model of the limited-availability group

In the considered model, m traffic classes that belong to the set $\mathbb{M} = \{1, 2, ..., m\}$ are defined. A given class c requires t_c BBUs to set up a new connection. The service time for class c calls has an exponential distribution with the parameter μ_c (service rate). In the group, the reservation mechanism has been applied. In accordance with the adopted reservation mechanism for a given class c, the reservation limit Q_c is introduced (Q_c is a certain occupancy state of the system, expressed in the number of BBUs being busy). The reservation mechanism can be applied to selected traffic classes from the set $\mathbb{M}$. The classes, in which the reservation limit has been introduced are grouped into a new set of classes $\mathbb{R}$, which is a sub-set of the set $\mathbb{M}$. The parameter R_c that determines the reservation area (a certain number of occupancy states of the system) has also been defined. This parameter can be expressed by the following formula:

$$R_c = V_F - Q_c. \tag{1}$$

The system admits a call of class c that belongs to the set $\mathbb{R}$ for service only when this call can be entirely carried by the resources of an arbitrary single link and when the number of free BBUs in the group is higher or equal to the value of the reservation area R_c. A call of class c that does not belong to the set $\mathbb{R}$ can be serviced when this call can be entirely carried by the resources of an arbitrary single link. This is, thus, an example of a system with a state-dependent service process, in which the state dependence is the result of the structure of the group and the introduced reservation mechanism. An example of the limited-availability group with reservation mechanism applied for only class 1 is presented in Figure 2.

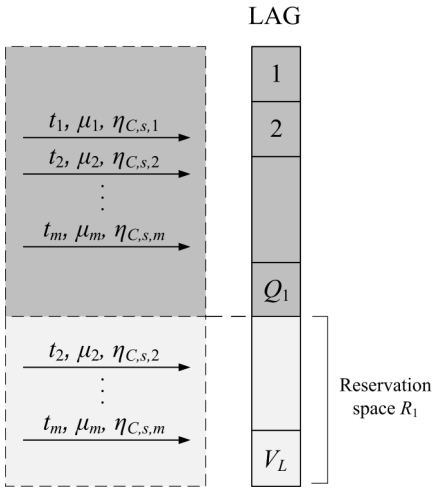


Figure 2. Model of the limited-availability group with reservation mechanisms

The group is offered three types of Erlang (Poisson call streams), Engset (binomial call streams) and Pascal (negative binomial call streams) traffic streams [13]. The selected types of traffic cover three different types of the dependence between the mean arrival rates of calls and the occupancy state of the system: (1) the mean arrival rate of new calls does not depend on the occupancy state of the system (Erlang traffic), (2) the mean arrival rate of new calls decreases with the increase in the occupancy state of the system (Engset traffic), (3) the mean arrival rate of new calls increases with the increase in the occupancy state of the system.

Each traffic stream is generated by sources that belong to the corresponding set of traffic sources $\mathbb{Z}_{C,s}$. In set $\mathbb{Z}_{C,s}$ index C denotes the type of traffic stream generated by sources, which belong to this set and takes the value I for Erlang traffic stream, J for Engset traffic stream and K for Pascal traffic stream, respectively, while index s denotes the number of the set, the sources of which generate a given type of traffic stream. In the system, the s_I sets of traffic sources that generate Erlang traffic streams are defined, as well as s_J sets of traffic sources that generate Engset traffic streams and s_K sets of traffic sources that generate Pascal traffic streams. The total number of the sets of traffic sources is $S = s_I + s_J + s_K$. The sources that belong to the set $\mathbb{Z}_{C,s}$ can generate calls from the set $\mathbb{C}_{C,s} = \{1, 2, ..., c_{C,s}\}$ of traffic classes according to the available set of services.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

The participation of class c (from the set $\mathbb{M}$) in the traffic structure of traffic generated by sources from the set $\mathbb{Z}_{C,s}$ is determined by the parameter $\eta_{C,s,c}$, which, for particular sets of Erlang, Engset and Pascal traffic sources, satisfies the following dependencies:

$$\sum_{c=1}^{c_{I,i}} \eta_{I,i,c} = 1, \qquad \sum_{c=1}^{c_{J,j}} \eta_{J,j,c} = 1, \qquad \sum_{c=1}^{c_{K,k}} \eta_{K,k,c} = 1. \tag{2}$$

To determine the value of traffic $A_{I,i,c}$ offered by Erlang sources that belong to the set $\mathbb{Z}_{I,i}$ as well as the traffic value $A_{J,j,c}(n)$ offered by Engset sources from the set $\mathbb{Z}_{J,j}$ and traffic $A_{K,k,c}(n)$ offered by Pascal sources from the set $\mathbb{Z}_{K,k}$ that generate calls of class c in the state of n busy BBUs, we use the following formulas [12]:

$$A_{I,i,c} = \eta_{I,i,c}\lambda_{I,i}/\mu_c. \tag{3}$$

$$A_{J,j,c}(n) = \eta_{J,j,c}N_{J,j}\alpha_{J,j}\sigma_{J,j,c,T}(n), \tag{4}$$

$$A_{K,k,c}(n) = \eta_{K,k,c}S_{K,k}\beta_{K,k}\sigma_{K,k,c,T}(n), \tag{5}$$

$$\sigma_{J,j,c,T}(n) = [\eta_{J,j,c}N_{J,j} - y_{J,j,c}(n)]/\eta_{J,j,c}N_{J,j}, \tag{6}$$

$$\sigma_{K,k,c,T}(n) = [\eta_{K,k,c}S_{K,k} + y_{K,k,c}(n)]/\eta_{K,k,c}S_{K,k}, \tag{7}$$

where:

- $\lambda_{I,i}$ – the mean arrival rate of new calls generated by a single Poisson source that belongs to the set $\mathbb{Z}_{I,i}$
- $\eta_{J,j,c}$ – the parameter that determines the participation of calls of class c in traffic generated by sources that belong to the set $\mathbb{Z}_{J,j}$,
- $\eta_{K,k,c}$ – the parameter that determines the participation of calls of class c in traffic generated by sources that belong to the set $\mathbb{Z}_{K,k}$,
- $N_{J,j}$ – the number of Engset traffic sources that belong to the set $\mathbb{Z}_{J,j}$,
- $S_{K,k}$ – the number of Pascal traffic sources that belong to the set $\mathbb{Z}_{K,k}$,
- $y_{J,j,c}(n)$ – the average number of calls of class c generated by Engset sources that belong to the set $\mathbb{Z}_{J,j}$ currently serviced in the system in the occupancy state n,
- $y_{K,k,c}(n)$ – the average number of calls of class c generated by Pascal sources that belong to the set $\mathbb{Z}_{K,k}$ currently serviced in the system in the occupancy state n,
- $\alpha_{J,j}$ – the average traffic intensity of traffic generated by a single Engset source that belongs to the set $\mathbb{Z}_{J,j}$, determined by the following formula:

$$\alpha_{J,j} = \sum_{c=1}^{c_{J,j}} \eta_{J,j,c}\frac{\gamma_{J,j}}{\mu_c}, \tag{8}$$

  where $\gamma_{J,j}$ – the mean arrival rate of new calls generated by a single Engset source that belongs to the set $\mathbb{Z}_{J,j}$,

- $\beta_{K,k}$ – the average traffic intensity of traffic generated by a single Pascal source that belongs to the set $\mathbb{Z}_{K,k}$, defined by the following formula:

$$\beta_{K,k} = \sum_{c=1}^{c_{K,k}} \eta_{K,k,c}\frac{\gamma_{K,k}}{\mu_c}, \tag{9}$$

  where $\gamma_{K,k}$ – the mean arrival rate of new calls generated by a single Pascal source that belongs to the set $\mathbb{Z}_{K,k}$.

We can notice that – according to Formulas (6) and (7) – with the case of Engset sources, the mean arrival rate of new calls of individual traffic classes decreases with the increase in the occupancy state of the system, whereas in the case of Pascal sources the mean arrival rate of new calls of individual traffic classes increases with the increase in the occupancy state of the system.

Taking into consideration the influence of the specific structure of the limited-availability group on the process of a determination of the occupancy distribution using Kaufman-Roberts recursion, in [14] the use of conditional coefficients of passing $\sigma_{c,S}(n)$, was proposed. The value of the parameter $\sigma_{c,S}(n)$ does not depend on incoming call process and can be determined as follows [14]:

$$\sigma_{c,S}(n) = \frac{F(V_L - n, \upsilon, f, 0) - F(V_L - n, \upsilon, t_c - 1, 0)}{F(V_L - n, \upsilon, f, 0)}, \tag{10}$$

where $F(x, \upsilon, f, t)$ is the number of arrangement of x free BBUs in υ links, calculated with the assumption that capacity of each link is equal to f BBUs and each link has at least t free BBUs:

$$F(x, \upsilon, f, t) = \sum_{r=0}^{\left\lfloor \frac{x-\upsilon t}{f-t+1} \right\rfloor} (-1)^r \binom{\upsilon}{r} \binom{x - \upsilon(t-1) - 1 - r(f-t+1)}{\upsilon - 1}. \tag{11}$$

Observe that in the case of the considered model of the limited-availability group with multi-service traffic sources and reservation the operation of the reservation mechanism introduces an additional dependence between the service stream in the system and the current state of the system. To determine this dependence, the parameter $\sigma_{c,R}(n)$ is introduced. The parameter $\sigma_{c,R}(n)$ can be calculated using the following formula:

$$\sigma_{c,R}(n) = \begin{cases} 1 & \text{for } n \leqslant Q_c \wedge c \in \mathbb{R}, \\ 0 & \text{for } n > Q_c \wedge c \in \mathbb{R}, \\ 1 & \text{for } c \notin \mathbb{R}. \end{cases} \tag{12}$$

The reservation mechanism is introduced to the group regardless of its structure, which allows us to carry on with product-form determination of the total coefficient of passing

(transition coefficient) $\sigma_{c,\text{Tot}}(n)$ in the limited-availability group:

$$\sigma_{c,\text{Tot}}(n) = \sigma_{c,S}(n) \cdot \sigma_{c,R}(n). \quad (13)$$

Having the values of offered traffic $A_{I,i,c}$, $A_{J,j,c}(n)$, $A_{K,k,c}(n)$ and the total coefficient of passing $\sigma_{c,\text{Tot}}(n)$ at our disposal, we are in position to modify the original Kaufman-Roberts formula [15] [16] in order to determine the occupancy distribution in the limited-availability group with multi-service traffic sources and the reservation mechanism:

$$\begin{aligned} n[P_n]_{V_L} &= \sum_{i=1}^{s_I}\sum_{c=1}^{c_{I,i}} A_{I,i,c}\sigma_{c,\text{Tot}}(n-t_c)t_c[P_{n-t_c}]_{V_L} + \\ &+ \sum_{j=1}^{s_J}\sum_{c=1}^{c_{J,j}} A_{J,j,c}(n-t_c)\sigma_{c,\text{Tot}}(n-t_c)t_c[P_{n-t_c}]_{V_L} + \\ &+ \sum_{k=1}^{s_K}\sum_{c=1}^{c_{K,k}} A_{K,k,c}(n-t_c)\sigma_{c,\text{Tot}}(n-t_c)t_c[P_{n-t_c}]_{V_L}, \quad (14) \end{aligned}$$

where $[P_n]_{V_L}$ is the occupancy distribution (the probability of n busy BBUs) in a system with the capacity V_L, and the parameter $\sigma_{c,\text{Tot}}(n)$ determines the additional dependence between the service stream and the current state of the system resulting from the specific structure of the group and the applied reservation mechanism.

Having the values of individual state probabilities $[P_n]_{V_L}$, determined on the basis of Formula (14), we are in position to determine the average number of serviced calls of class c, generated by sources that belong to the sets $\mathbb{Z}_{J,j}$ (Engset) and $\mathbb{Z}_{K,k}$ (Pascal sources). For this purpose, we use the following formulas:

$$y_{J,j,c}(n) = \begin{cases} A_{J,j,c}(n-t_c)\sigma_{c,\text{Tot}}(n-t_c)[P_{n-t_c}]_{V_L}/[P_n]_{V_L} \\ \qquad \text{for } n \leqslant V_L, \\ 0, \qquad \text{for } n > V_L. \end{cases} \quad (15)$$

$$y_{K,k,c}(n) = \begin{cases} A_{K,k,c}(n-t_c)\sigma_{c,\text{Tot}}(n-t_c)[P_{n-t_c}]_{V_L}/[P_n]_{V_L} \\ \qquad \text{for } n \leqslant V_L, \\ 0, \qquad \text{for } n > V_L. \end{cases} \quad (16)$$

The knowledge of the occupancy $[P_n]_{V_L}$ is required to determine the parameters $y_{J,j,c}(n)$ and $y_{K,k,c}(n)$. Whereas, to determine the occupancy $[P_n]_{V_L}$ it is necessary to know the values of the parameters $y_{J,j,c}(n)$ and $y_{K,k,c}(n)$. Equations (15), (16) and (14) form thus a set of confounding equations. To solve a given set of confounding equations it is necessary to employ iterative methods [17] [18].

Assuming that the distribution $[P_n^{(l)}]_{V_L}$ is the occupancy distribution, determined in the l-th iteration, while $y_{J,j,c}^{(l)}(n)$ and $y_{K,k,c}^{(l)}(n)$ define the average number of serviced calls of class c generated by traffic sources that belong respectively to the sets $\mathbb{Z}_{J,j}$ and $\mathbb{Z}_{K,k}$, we can write:

$$y_{J,j,c}^{(l+1)}(n) = \begin{cases} A_{J,j,c}^{(l)}(n-t_c)\sigma_{c,\text{Tot}}(n-t_c)[P_{n-t_c}^{(l)}]_{V_L}/[P_n^{(l)}]_{V_L} \\ \qquad \text{for } n \leqslant V_L, \\ 0, \qquad \text{for } n > V_L. \end{cases} \quad (17)$$

$$y_{K,k,c}^{(l+1)}(n) = \begin{cases} A_{K,k,c}^{(l)}(n-t_c)\sigma_{c,\text{Tot}}(n-t_c)[P_{n-t_c}^{(l)}]_{V_L}/[P_n^{(l)}]_{V_L} \\ \qquad \text{for } n \leqslant V_L, \\ 0, \qquad \text{for } n > V_L. \end{cases} \quad (18)$$

The iteration process, involving Formulas (14), (17) and (18), terminates when the assumed accuracy ϵ of the iteration process is reached:

$$\forall_{0\leq n\leq V} \left| \frac{y_{J,j,c}^{l-1}(n) - y_{J,j,c}^{(l)}(n)}{y_{J,j,c}^{(l)}(n)} \right| \leqslant \epsilon, \quad (19)$$

$$\forall_{0\leq n\leq V} \left| \frac{y_{K,k,c}^{l-1}(n) - y_{K,k,c}^{(l)}(n)}{y_{K,k,c}^{(l)}(n)} \right| \leqslant \epsilon. \quad (20)$$

Subsequently, we are in position to determine the blocking probability for calls of class c that belong to the set $\mathbb{M} = \{1, 2, ..., m\}$:

$$E_c = \sum_{n=0}^{V_L} [P_n]_{V_L} \left[1 - \sigma_{c,\text{Tot}}(n)\right]. \quad (21)$$

III. Numerical Results

The presented method for a determination of the blocking probability in systems with multi-service traffic sources and reservation mechanisms is an approximate method. In order to confirm adopted assumptions, the results of the analytical calculations were compared with the simulation data. The research was carried for three systems, which are described below:

1) Limited-availability system No. 1
   - Capacity: $\upsilon = 2$, $f = 20$ BBUs, $V_L = 40$ BBUs,
   - Number of traffic classes: 3
   - Structure of traffic: $t_1 = 1$ BBU, $\mu_1^{-1} = 1$, $t_2 = 2$ BBUs, $\mu_2^{-1} = 1$,$t_3 = 6$ BBUs, $\mu_3^{-1} = 1$, $R_1 = R_2 = 33$ BBUs
   - Sets of sources: $\mathbb{C}_{I,1} = \{1, 2\}$, $\eta_{I,1,1} = 0.6$, $\eta_{I,1,2} = 0.4$, $\mathbb{C}_{J,2} = \{2, 3\}$, $\eta_{J,2,2} = 0.7$, $\eta_{J,2,3} = 0.3$, $N_2 = 60$
2) Limited-availability system No. 2
   - Capacity: $\upsilon = 2$, $f = 30$ BBUs, $V_L = 60$ BBUs,
   - Number of traffic classes: 3
   - Structure of traffic: $t_1 = 1$ BBU, $\mu_1^{-1} = 1$, $t_2 = 3$ BBUs, $\mu_2^{-1} = 1$,$t_3 = 7$ BBUs, $\mu_3^{-1} = 1$, $R_1 = R_2 = 51$ BBUs
   - Sets of sources: $\mathbb{C}_{I,1} = \{1\}$, $\eta_{I,1,1} = 1.0$, $\mathbb{C}_{J,2} = \{1, 2\}$, $\eta_{J,2,1} = 0.6$, $\eta_{J,2,2} = 0.4$, $N_2 = 50$,

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

$\mathbb{C}_{K,3} = \{2,3\}$, $\eta_{K,3,2} = 0.7$, $\eta_{K,3,3} = 0.3$, $S_3 = 50$

3) Limited-availability system No. 3

- Capacity: $\upsilon = 4$, $f = 20$ BBUs, $V_L = 80$ BBUs,
- Number of traffic classes: 4
- Structure of traffic: $t_1 = 1$ BBU, $\mu_1^{-1} = 1$, $t_2 = 2$ BBUs, $\mu_2^{-1} = 1$,$t_3 = 4$ BBUs, $\mu_3^{-1} = 1$, $t_4 = 9$ BBUs, $\mu_4^{-1} = 1$, $R_1 = R_2 = R_3 = 63$ BBUs
- Sets of sources: $\mathbb{C}_{I,1} = \{1,2\}$, $\eta_{I,1,1} = 0.6$, $\eta_{I,1,2} = 0.4$, $\mathbb{C}_{J,2} = \{2,3\}$, $\eta_{J,2,2} = 0.7$, $\eta_{J,2,3} = 0.3$, $N_2 = 70$, $\mathbb{C}_{K,3} = \{2,3,4\}$, $\eta_{K,3,2} = 0.3$, $\eta_{K,3,3} = 0.2$, $\eta_{K,3,4} = 0.5$, $S_3 = 140$

The results of the research study are presented in Figures 3-5, depending on the value of traffic a offered to a single BBU. The mean value of offered traffic a can be calculated using following equation:

$$a = \left[\sum_{i=1}^{s_I} \lambda_{I,i} \sum_{c=1}^{c_{I,i}} t_c \eta_c \sum_{c=1}^{c_{I,i}} \eta_c / \mu_c + \right.$$
$$+ \sum_{j=1}^{s_J} \gamma_{J,j} N_{J,j} \sum_{c=1}^{c_{J,j}} t_c \eta_c \sum_{c=1}^{c_{J,j}} \eta_c / \mu_c +$$
$$\left. + \sum_{k=1}^{s_K} \gamma_{K,k} S_{K,k} \sum_{c=1}^{c_{K,k}} t_c \eta_c \sum_{c=1}^{c_{K,k}} \eta_c / \mu_c \right] / V_L. \quad (22)$$

The results of the simulation are shown in the charts in the form of marks with 95% confidence intervals that have been calculated according to the t-Student distribution for the five series with 1,000,000 calls of each class. For each of the points of the simulation, the value of the confidence interval is at least one order lower than the mean value of the results of the simulation. In many a case, the value of the simulation interval is lower than the height of the sign used to indicate the value of the simulation experiment.

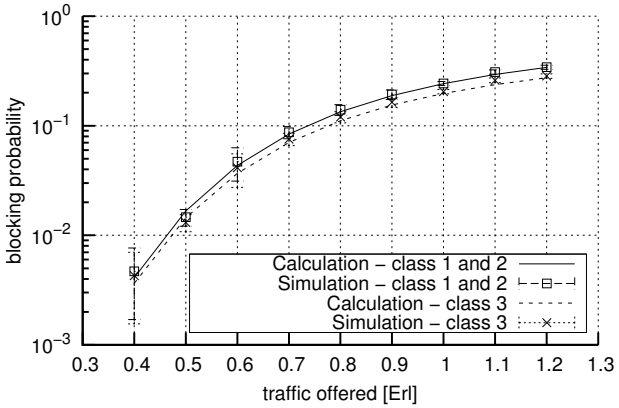


Figure 3. Blocking probability in the limited-availability group No. 1 with reservation mechanism; the reservation mechanism equalizes the blocking probability for calls of class 1 and 2

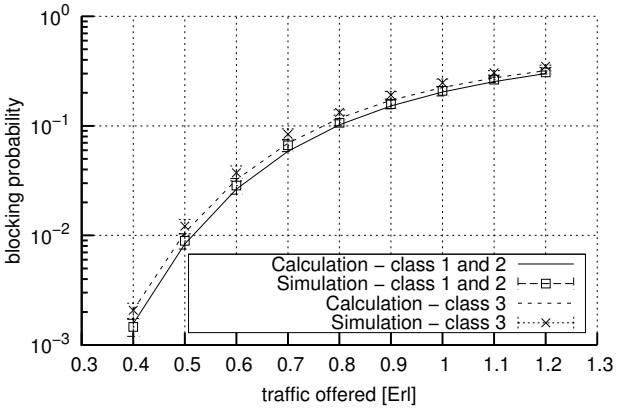


Figure 4. Blocking probability in the limited-availability system No. 2 with reservation mechanism; the reservation mechanism equalizes the blocking probability for calls of class 1 and 2

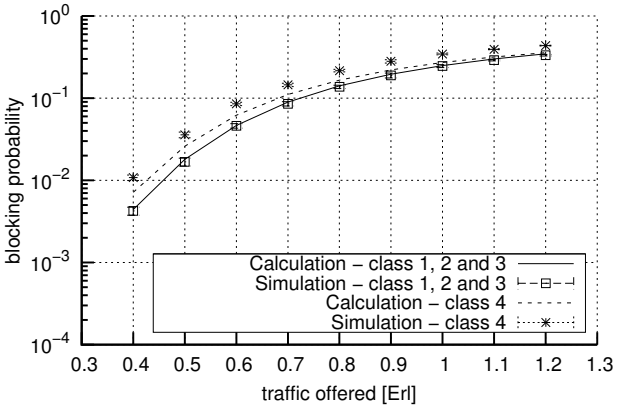


Figure 5. Blocking probability in the limited-availability system No. 3 with reservation mechanism; the reservation mechanism equalizes the blocking probability for calls of class 1, 2 and 3

IV. Conclusion and Further Work

This paper proposes a new method for a calculation of the occupancy distribution and the blocking probability in limited-availability systems with multi-service traffic sources and reservation mechanisms. The method can be used in modeling connection handoff between cells in cellular systems [8], as well as in modeling outgoing directions of switching networks [19]. The proposed method is based on the iterative algorithm for a determination of the average value of traffic sources being serviced in particular states of the system. The results of analytical calculations were compared with the simulation data, which confirmed high accuracy of the proposed method. The proposed method is not complicated and can be easily implemented.

In the further work we plan to develop analytical models of the multi-service systems with multi-service sources, in which different call admission control mechanisms will be applied, i.e., an analytical model of multi-service networks with threshold mechanisms and multi-service sources, and

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

a model of multi-service systems with hysteresis and multi-service sources.

REFERENCES

[1] 3GPP, "High Speed Packet Access (HSPA) evolution; Frequency Division Duplex (FDD)," 3rd Generation Partnership Project (3GPP), TR 25.999, Mar. 2008, <retrieved: April, 2012>. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/25999.htm

[2] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRAN); Overall description; Stage 2," 3rd Generation Partnership Project (3GPP), TS 36.300, Sep. 2008, <retrieved: April, 2012>. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/36300.htm

[3] ——, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," 3rd Generation Partnership Project (3GPP), TS 23.401, Sep. 2008, <retrieved: April, 2012>. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/23401.htm

[4] A. Hadden, "Mobile broadband - where the next generation leads us [industry perspectives]," *Wireless Communications, IEEE*, vol. 16, no. 6, pp. 6–9, Dec. 2009.

[5] S. Ortiz Jr., "4G wireless begins to take shape," *Computer*, vol. 40, no. 11, pp. 18–21, 2007.

[6] S. R. Subramanya, "Emerging mobile technologies and issues," in *Proceedings of the International Symposium on Collaborative Technologies and Systems*. Washington, DC, USA: IEEE Computer Society, 2006, p. 172.

[7] V. G. Vassilakis, I. D. Moscholios, and M. D. Logothetis, "Call-level performance modelling of elastic and adaptive service-classes with finite population," *IEICE Transactions on Communications*, vol. E91-B, no. 1, pp. 151–163, 2008.

[8] M. Głąbowski, M. Sobieraj, and M. Stasiak, "Analytical modeling of the influence of connection handoff on UMTS traffic characteristics," in *Proceedings of Third Advanced International Conference on Telecommunications*, Morne, may 2007.

[9] 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. E95-B, no. 1, pp. 120–132, 2012.

[10] M. Głąbowski, M. Sobieraj, and M. Stasiak, "Evaluation of traffic characteristics of UMTS with bandwidth reservation and handoff mechanism," in *Proceedings of 14th IEEE International Conference On Telecommunications*, Penang, May 2007, pp. 736–740.

[11] ——, "Blocking probability calculation in UMTS networks with bandwidth reservation, handoff mechanism and finite source population," in *Proceedings of 7th International Symposium on Communications and Information Technologies*, Sydney, Oct. 2007, pp. 433–438.

[12] ——, "Analytical modeling of multi-service systems with multi-service sources," in *Proceedings of 16th Asia-Pacific Conference on Communications (APCC)*. Auckland, New Zealand: IEEE, Oct. 2010, pp. 285–290.

[13] H. Akimaru and K. Kawashima, *Teletraffic: Theory and Application*. Berlin-Heidelberg-New York: Springer, 1999.

[14] M. Stasiak, "Blocking probability in a limited-availability group carrying mixture of different multichannel traffic streams," *Annales des Télécommunications*, vol. 48, no. 1-2, pp. 71–76, 1993.

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

[16] J. Roberts, "A service system with heterogeneous user requirements — application to multi-service telecommunications systems," in *Proceedings of Performance of Data Communications Systems and their Applications*, G. Pujolle, Ed. Amsterdam: North Holland, 1981, pp. 423–431.

[17] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Modeling product-form state-dependent systems with bpp traffic," *Perform. Eval.*, vol. 67, pp. 174–197, March 2010.

[18] M. Głąbowski, M. Stasiak, and J. Weissenberg, "Properties of recurrent equations for the full-availability group with bpp traffic," *Mathematical Problems in Engineering*, vol. 2012, p. 17, 2012, article ID 547909.

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

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

An Evaluation of IPv6 in Simulation using OPNET Modeler

Brittany Clore*†, Matthew Dunlop*†, Randolph Marchany†, Joseph Tront*
*Bradley Department of Electrical and Computer Engineering
†Virginia Tech Information Technology Security Office
Virginia Tech
Blacksburg, Virginia 24061, USA
e-mail: {clore, dunlop, marchany, jgtront}@vt.edu

***Abstract*— Simulation is vital to be able to test various network topologies and new components in a cost effective manner. With the push to adopt Internet Protocol version 6 (IPv6), many network administrators need to be able to test their hardware and specialized applications before deploying them on a live network. OPNET Modeler provides the capability to simulate an IPv6 network and the OPNET System in the Loop, an add-on module, allows for real devices to be tested over the simulated network. This study evaluates the support of IPv6 in OPNET Modeler 16.1 with the System in the Loop module. The results show that this module does not fully support IPv6 at this time but with improvements can be an important part to planning and implementing IPv6 networks.**

Keywords - Simulation; IPv6; System in the Loop; OPNET

I. INTRODUCTION

With the assignment of the last block of IPv4 addresses in February 2011, IPv6 is being pushed to rapidly be deployed in new networks. These networks can have a wide range of devices connected into it including specialized software. Before committing to implement a full scale IPv6 production network, simulation of the environment allows network administrators to analyze how their configuration will function. There is a wide range of simulation tools available that can achieve this goal. OPNET Modeler is a commercial solution that provides a wide range of simulated network devices from workstations to switches and routers. While users primarily interact within the graphical user interface, the software is expandable with user written code. The code is C based, with OPNET providing its own classes and functions [1].

While simulating a basic network is vital to examine for IPv6 readiness, many software and hardware vendors are adapting their technologies to support IPv6. There is a need to be able to test these products on an IPv6 network. Simulation is a cost effective way to conduct testing due the capability to simulate various network topologies, sizes, and conditions [2]. The problem is that there are very few network modeling tools that are IPv6 capable. The main simulators that claim to be IPv6 capable are NS3 [3], OMNeT++ [4], and OPNET [5]. Of these, OPNET possesses the best capability to tie in live systems to a simulation environment. For that reason, OPNET was selected as the network simulator to test IPv6 research in the Virginia Tech Information Technology Security Office. OPNET's System in the Loop module is a way to test actual products on an IPv6 network without having to convert the code into simulation. This module was the focus of the study to test the extent of the IPv6 support.

This remainder of this paper is organized as follows: Section II describes other work related to simulating IPv6 in OPNET. Section III provides some background on the OPNET System in the Loop module as well as IPv6 in general. Section IV discusses the set up considerations that are needed for proper functionality. Section V describes the design of the study while Section VI demonstrates the results. The paper is concluded in Section VII along with a discussion of some future work.

II. RELATED WORK

OPNET Modeler is a widely used simulation program that advertises IPv6 support. There have been various studies that have assessed applications in IPv6 within a fully simulated network. One study by Aziz et al. looked at the performance of video and voice traffic in IPv6 [6]. The authors used OPNET Modeler to run simulations in IPv4 and IPv6 to compare throughput and were able to show that IPv6 slightly decreases throughput due to its packet overhead. Le et al. [7] assessed the Mobile IPv6 model for IPv6 header support and routing. They found that the model was able to correctly handle IPv6. These both show that Modeler is capable of simulating an IPv6 network successfully.

Green et al. [8] characterized a test bed for IPv6 applications. Their setup was a simulated network communicating between one real device using a "hardware in the loop" scheme which is very similar to System in the Loop. The difference is that System in the Loop can test a single software piece without specialized hardware. One of their observations was that OPNET does not provide the capability to do a one-to-one match with real packet data to simulated packet data but this could be added with additional code. Their scheme did not function for real time traffic but rather worked for a single non-real time stream of traffic.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

III. BACKGROUND

OPNET Modeler is a tool that allows for a wide range of simulation. To extend the simulation, modules can be added on that add extra features. One module is the System in the Loop module. It is also important to understand IPv6 to identify what options need to be implemented to verify that this module has IPv6 support.

A. OPNET System in the Loop

OPNET's System in the Loop allows for communication between real, physical devices or software and a simulated network. It does this by using a specialized node that listens on a given network interface and filters incoming packets (real to simulated network) using the Berkeley Packet Filter syntax. Once it receives the packets, a translation function converts the packet headers and payload into the simulation packet format. The module currently supports the translation of the following protocols: IPv4, IPv6, ICMP, ICMPv6, OSPF, RIPv1, RIPv2, TCP, UDP, and FTP [1].

There are three configurations in which the simulation communication can be set up: physical device to simulated device, simulated device to simulated device through a real device and real device to real device through a simulated device. The assessment was done for the third type of communication setup, real device to real device through simulation. This configuration provided the capability to evaluate packet behavior as real packets enter the simulation environment and again as the same packets are translated back into real packets for delivery to a live destination.

B. IPv6 Background

IPv6 differs significantly from Internet Protocol version 4 (IPv4). The most noticeable difference is that IPv6 uses a 128 bit addressing space while IPv4 uses a 32 bit addressing space [9]. Within the OPNET simulation code this translates into using a pointer versus using a defined type such as a double. To provide an idea of the scope of the IPv6 address size, the entire IPv4 address space fits into a single IPv6 subnet over four billion times.

Another difference is the packet header. Where IPv4 headers were of variable length due to the possible inclusion of options, IPv6 headers are a fixed 40 bytes [9]. Options are includes as extension headers and become part of the payload. Extension headers do not have a specified order and contain a next header field that acts like a chain within the header to connect all the extension headers. Extension headers are used to specify what protocol is next being used in the packet as well as other functions like fragmentation and Internet Protocol Security (IPSec) options. An extension header also exists where users can define their own functionality. This type of extension header is referred to a Destination Options header. Destination options contain information that only pertains to the final intended recipient. This flexibility poses a problem in the packet translation to simulation.

Another significant difference is how IPv6 accomplishes address resolution. Due to the large address size, hosts in IPv6 generate their own addresses. This reduces the management burden placed on network managers. Hosts use a process called Stateless Address Auto configuration (SLAAC) to generate addresses. SLAAC addresses are advertised to other network hosts using the Neighbor Discovery Protocol (NDP), which replaces the address resolution protocol (ARP) used in IPv4. NDP uses a series of Internet Control Messaging Protocol version 6 (ICMPv6) messages to advertise addresses as well as solicit for router and other hosts. NDP removes the need to perform certain tasks like specifying a gateway, as this is accomplished by router solicitations and advertisements by the protocol. In addition to NDP messages, ICMPv6 includes other error message types also used by ICMP in IPv4.

IV. SET UP

There are specific configuration details that are required for OPNET's System in the Loop module to operate properly. For example, it is important to signify the right interface by including the source Ethernet media access control (MAC) address in the packet filter. Other filters for protocol can be used to further limit the traffic that the module translates into simulation.

The simplest interface configuration is to have one physical network interface per real device. This is not always feasible and so it is possible to have one interface handling the traffic of all real devices; however, the packet filter has to be very specific otherwise traffic can be sent through the wrong section of the simulated network.

Within the simulation environment, a System in the Loop node can only be connected to another node through Ethernet. This connection is defined as a duplex 10Gbps link. Half-duplex links are not allowed. Also within the simulation environment, the System in the Loop node has to define the translation function it's using as well. For this study, the default translation function was being assessed.

V. DESIGN

In its current form, OPNET's System in the Loop supports a small set of protocols. The purpose of this study was to assess the support of IPv6. The main goal was to achieve communication through the simulated network with the intent to measure various IPv6 applications performances.

The design for this study was an isolated network in which two physical nodes were connected by a simulated set of routers. The physical nodes were virtual machines that were hosted on the same machine that runs OPNET Modeler. The virtual machines' network interfaces were bridged with two separate network interface cards that were installed on the host machine. These cards solely handle traffic to and from the virtual machines. Fig. 1 depicts the layout of the virtual machines and simulation. The encompassing box represents the host machine.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

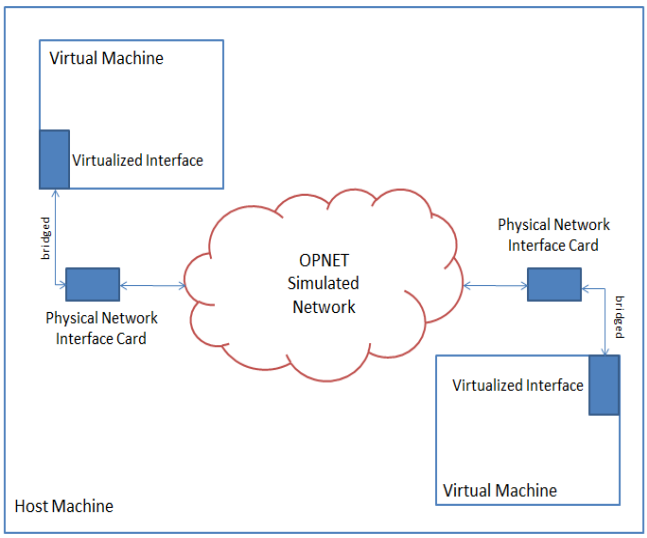


Figure 1. Layout of Virtual Machines and OPNET Simulated Network on the Host Machine

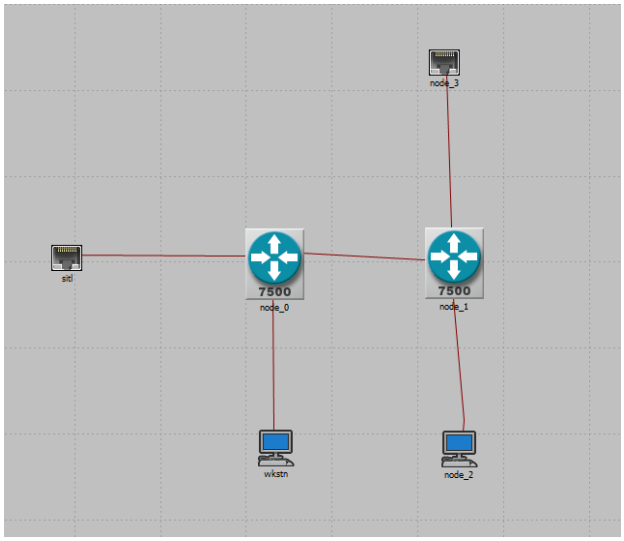


Figure 2. OPNET Simulated Network

The System in the Loop module listens on these network interfaces. The packet filter is defined to filter for IPv6 traffic coming from the MAC address of each virtual machine. The simulated network contains two routers and two workstations. Fig. 2 shows the topology of the simulated network. The workstation nodes are generic nodes defined by OPNET. The routers are simulated Cisco 7507 devices. The simulated workstations provided the option to test communication within simulation. The System in the Loop nodes are located at the right and to the top left. The icon for that node is an Ethernet port. To communicate, these nodes would have to make two hops. Both virtual machines run the Ubuntu 11.4 operating system which supports IPv6 networking.

Two categories of tests were run on the simulation. The first set was to achieve communication between the physical nodes with a set of ping messages. This tested connectionless ICMPv6 ping message support and Neighbor Discovery Protocol support. A standard 1-second ping was used as well as executing a 10,000 packet ping flood. The goal of the second set of tests was to test connection-oriented transmission control protocol (TCP) and hypertext transfer protocol (HTTP) communication using IPv6 addressing. This was achieved through accessing a webpage being hosted on one of the virtual machines and doing a series of files transfers using wget, a free software package that allows files transfer through the HTTP protocol. The file sizes transferred ranged from 1-kilobyte files to 1-gigabyte files. These tests were executed on the live nodes through the command line.

VI. RESULTS

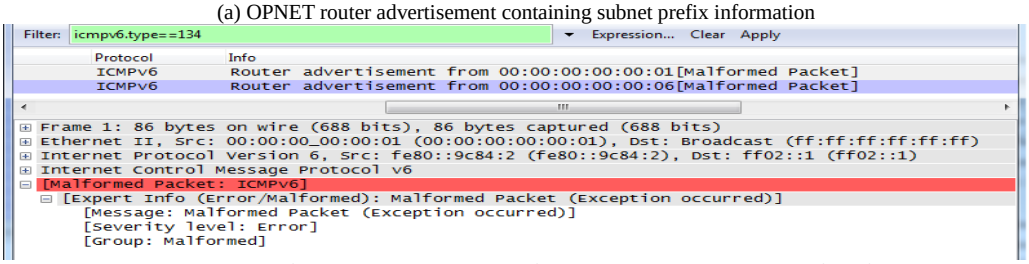
OPNET's System in the Loop proved to not fully support IPv6. There are two main issues that were found preventing this software module from being fully able to simulate real IPv6 communication. One issue related to proper support of NDP while the other issue was caused by lack of support for some ICMPv6 message types.

The first issue was caused by the inability to properly process NDP router advertisement messages. In IPv6, hosts rely on router advertisements sent by local routers to auto-configure addresses and to learn of possible gateways. Without these router advertisements, hosts cannot learn what subnet they are connected to nor which router is their closest gateway. OPNET is sending out router advertisements, but they are malformed. As a result, real systems connected to OPNET still need to statically set addresses and gateways. Fig. 3(a) shows the OPNET format for the router advertisement. The router advertisement contains a subnet prefix value and length. OPNET is unable to properly translate the packet. Fig 3(b) shows the corresponding packet in Wireshark as being malformed. In OPNET, there is a way to manually set the prefix, but this is also not translated in a correct manner. It is not clear whether the router advertisement is malformed within the simulation and the simulated nodes understand the bad packet or if the packet becomes malformed due to processing by OPNET's System in the Loop module.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

```
* Packet Fields:
      Index Name          Size I Type             Value
      ----- ------------ ------ - ---------------- --------
          0 fields          448   structure        0x02F0C598
                  Type              int          Rtr Advertise    (8)
                  Rtr Lifetime      int          1800            (16)
                  Src LL Addr       int          0   (80)
                  Prefixes          structure                    (256)
                                   2005:0:0:2:0:0:0:0/64 (on-link)
          1 echo_packet       0    packet           -
```

(a) OPNET router advertisement containing subnet prefix information



(b) Wireshark capture of the resulting translated malformed packet with no subnet prefix information

Figure 3. Comparison of OPNET router advertisement with the translated packet in Wireshark

The second issue is that OPNET only accepts a small set of ICMPv6 message types and does so improperly. When first trying to communicate with ping messages (the first category of tests) between physical machines, the simulation would fail because it did not recognize the ICMPv6 message type. However, ping request and response message types are in the OPNET supported set. Further investigation in the code revealed that the top bytes were getting improperly set. This error could be caused by two things. One cause is the use of an improper type to handle the information. The second cause is that the translation function from real to simulated packet adds erroneous data. The fix for this issue was to mask the lower two bytes of the message type field in the simulation code for the IPv6 Neighbor Discovery process node. The specific code for this is:

icmp_pk_fields_ptr->message_type & 0xff.

The resulting value is a correct message type. Fig. 5 shows the simulation error message before making this fix.

Two smaller concerns are that the current translation functions do not support many IPv6 extension headers or any IPv6 routing protocols. These issues were discovered in the second category of tests. These concerns were known before doing the assessment from training classes provided by OPNET [10]. For full IPv6 support, extension headers are must because of many features inherent to IPv6. For example, IPv6 includes native support of IPSec. IPSec is implemented through extension headers, which are currently not supported by OPNET. Further investigation was done using Scapy [11], a packet manipulation tool. Packets using each extension header were created and sent through simulation. Fig. 6 shows an example of the hop by hop extension header packet being sent and OPNET's log message saying it is unsupported. It was found that no extension headers defined by RFC 2460 [9] were supported.

After fixing the message type and statically setting a gateway on the network interfaces, the tests did execute. Fig. 4 shows a graph of the IPv6 traffic received by the simulated router during the second category of tests. It clearly shows traffic is being translated into simulation. When the wget transfer finishes after the eight minute mark, the traffic received drops as expected. Both categories of tests produced similar graphs and results.

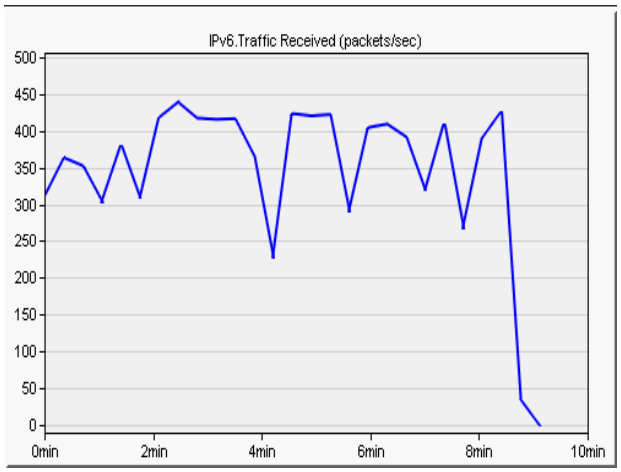


Figure 4. IPv6 traffic received by the simulated router

VII. Conclusion and Future Work

Out of the box, OPNET's System in the Loop is not yet ready to handle IPv6 simulation. Errors in handling essential fields of packets make it difficult to get the full potential out of the product. Complications that arise during the set up of network interface cards and virtual machines with System in the Loop can hinder the simulation and add an extra step to analyzing data.

Future work includes writing a new translation function for IPv6 packets that includes extension header support.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Further investigation into the malformed router advertisements needs to be done to see if a new translation function would solve this error. Using a different configuration to see the effect on translation delays is also planned.

With a translation function that supports the flexibility of IPv6 with its extension headers and fixes in the ICMPv6 message type handling, System in the Loop can be a viable simulation tool for network administrators.

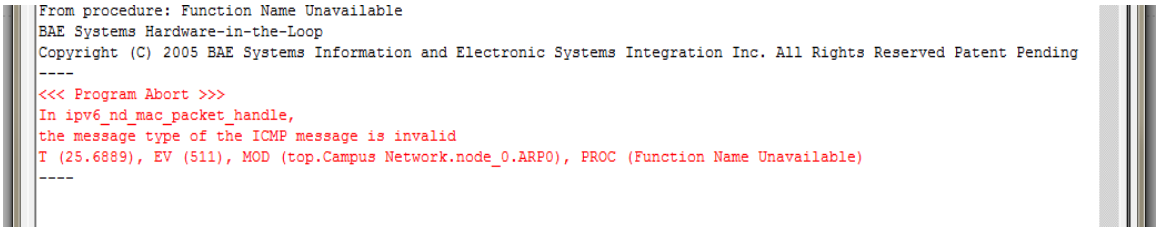
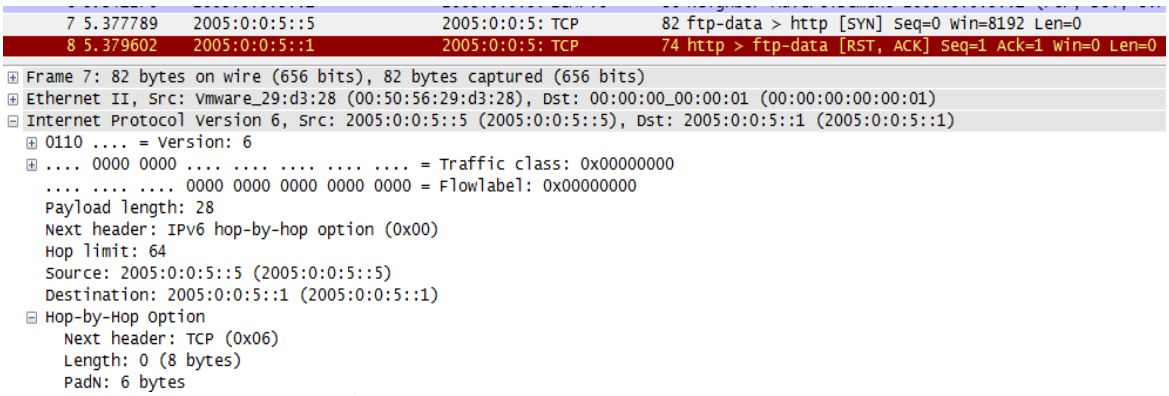


Figure 5. ICMP Message Type Error



(a) Wireshark capture of a packet using the hop by hop extension header being sent into simulation

(b) OPNET log output of unsupported packet

Figure 6. Attempt to send a packet using the hop by hop extension header into OPNET from a live machine

REFERENCES

[1] "OPNET Modeler" [Online] Available: http://www.opnet.com/solutions/network_rd/modeler.html, Accessed on 23 January 2012.

[2] Kaplan, G.; , "Simulating networks," Spectrum, IEEE , vol.38, no.1, pp.74-76, Jan 2001 doi: 10.1109/6.901148 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=901148&isnumber=19336G.

[3] "ns-3: IPv6 Class Reference" [Online] Available: http://www.nsnam.org/doxygen/classns3_1_1_ipv6.html, Accessed on 25 January 2012.

[4] "OMNeT++ IPv6 Suite" [Online] Available: http://www.omnetpp.org/omnetpp/doc_details/2137-ipv6suite, Accessed on 25 January 2012.

[5] "OPNET: IPv6 for R&D Specialized Model" [Online] Available: http://www.opnet.com/solutions/network_rd/simulation_model_library/ipv6.html, Accessed on 12 December, 2012.

[6] M. Aziz, M. Islam, and M. Khan, "Throughput Performance Evaluation of Video/Voice Traffic in IPv4/IPv6 Network," in Internation Journal of Computer Applications, vol. 35 no. 2, pp. 5-12, December 2011.

[7] D. Le, X. Fu, and D. Hogrefe, "Evaluation of Mobile IPv6 Based on an OPNET Model," unpublished.

[8] Green, D.; Mayo, R.; Ranga Reddy; , "IPv6 Application Performance Characterization Using a Virtual/Live Testbed," Military Communications Conference, 2006. MILCOM 2006. IEEE , vol., no., pp. 1-4, 23-25 Oct. 2006 doi: 10.1109/MILCOM.2006.302398 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4086641&isnumber=4043248

[9] S. Deering and R. Hinden, Internet Protocol, Version 6 (IPv6) Specification, IETF RFC 2460, December 1988; http://www.ietf.org/rfc/rfc2460.txt

[10] "OPNET Training" [Online] Available: https://www.opnet.com/training/index.html, Accessed on 23 January 2012.

[11] "Scapy" [Online] Available: http://trac.secdev.org/scapy, Accessed on 1 February 2012.

Techno-economic analysis for Rural Broadband Access Networks

Navneet Nayan, Rong Zhao, Nikolay Zhelev, Wolfgang Knospe
Detecon International GmbH, 53227 Bonn, Germany
Navneet.Nayan | Rong.Zhao | Nikolay.Zhelev | Wolfgang.Knospe@detecon.com

Carmen Mas Machuca
Technology University of Munich, 80290 München, Germany
cmas@tum.de

***Abstract*— This paper elaborates a techno-economic cost model for deploying Broadband Access Networks in rural areas around the world. It is aimed to come up with the major benefits and challenges associated with offering broadband access/services in rural areas and also to derive an effective solution towards this problem. The complete picture including all relevant factors impacting costs and benefits of rural broadband networks have been presented. A Technology Selection Strategy is also proposed to select the best-fit solution, subject to technical and economic modelling including regulation, revenue and funding. A quantitative analysis leading to an empirical techno-economic model for computing the total cost-benefits associated with rural broadband has been developed. A short insight into a Germany-based case study for rural broadband has also been depicted.**

Keywords - Rural Broadband Access; Techno-economic analysis; Cost model; Regulation; Funding

I. INTRODUCTION

The communication network market is widely accepted to be one of the most dynamic fields of business and technology. With a multitude of fast-paced technological innovations, ever-insatiable market demands and tightly coupled regulatory obligations; dynamism seems to exist in almost every aspect of this digitally networked environment.

The Internet has been a global connector, establishing links from one corner of the world to the other through a well defined globally distributed network. A series of innovations in this field of engineering propelled different technological adoptions in all segments, leading to higher capacity networks as we see it currently.

However, not much has driven the development of broadband communication networks for rural communities globally. This is primarily due to factors such as low return on investment potential for network operators, lower spending capability of rural populace, stringent regulatory landscapes and inadequate funding resources [1].

In order to understand the issue regarding rural broadband, and thereby, derive a feasible solution, it is vital to construct the overall picture of the components involved in the business case for rural networks. Most of the studies look for particular scenarios as for Africa [2] or India [3] although they do not include important aspects as regulation and funding.

This paper is organized as follows: Section II introduces the current state of rural areas as well as an overview of the benefits and major challenges involved in provisioning broadband services such as regulatory and funding aspects.. Section III revolves around the technology options for rural broadband services and their techno-economic benchmarking. An innovative approach to model the cost-benefits of rural broadband networks, including the relevant factors, has been described in Section IV. This is achieved using the Technology Selection Strategy elaborated in the same section. Section V describes the process of identifying the best fitting technical choice and its application to a German rural area.

II. RURAL BROADBAND ACCESS

A. Rural Broadband – Motivation

Rural broadband access aims to deliver efficient solutions to connect the rural (un- and underserved) community with access to the Internet at appropriate bandwidth. It is aimed to provide a previously un-served or underserved community with access to the Internet at a sufficient speed as to not be left behind or be disadvantaged from the subscribers in the city/localities with a proximity to regional/backbone networks. The connection will allow the subscribers to fast and efficiently use all services available on the Internet.

Multifarious e-services, triple play and specified services for rural areas are the main drivers to develop the rural broadband access. As illustrated in Fig. 1, broadband brings along tremendous amount of opportunities for growth and development. The direct and indirect effect of broadband Internet has been thoroughly researched and proposes substantial benefits in terms of economic growth [4, 5].

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

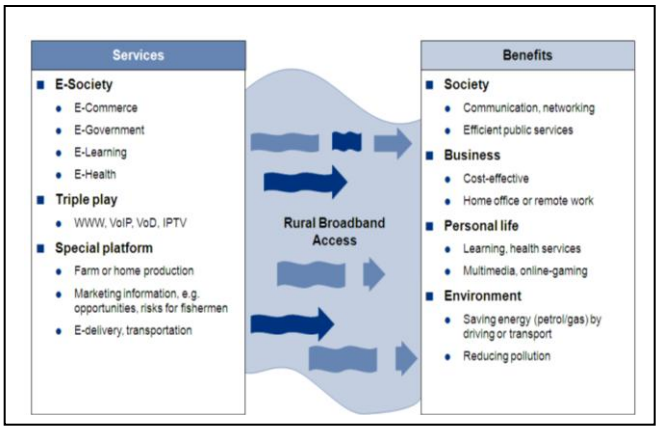


Figure 1. Services and Benefits of Rural Broadband

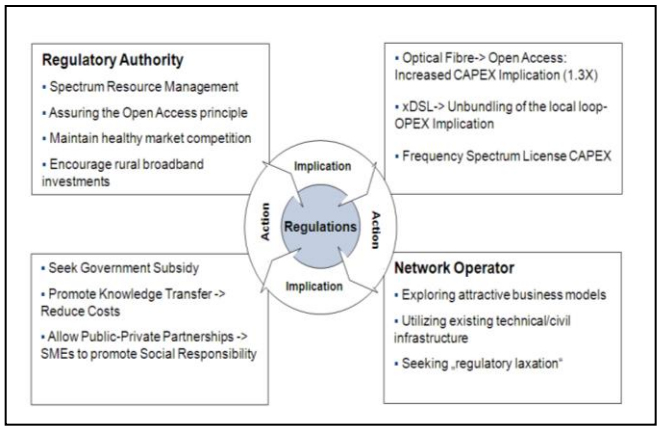


Figure 2. Regulatory Concerns of Rural Broadband

B. Rural Broadband – Challenges

Although the benefits of providing rural broadband access are countless, deploying rural networks face a number of challenges from multifarious domains. In general, the overall scope of factors hindering such initiatives can be summarized as follows:

Technical factors

A number of technical considerations must be taken into account while making the choice for a broadband technology.

- Distance and Topology: Technologies suited for special distance and topography
- Scalability: Technologies with expected bandwidth, and easily upgraded to future technology
- Resource Contention: reuse of existing infrastructure, shared or dedicated last mile platform
- Cost Efficiency: Technologies with low Capital and Operational Expenses (i.e. CAPEX and OPEX), e.g. lower energy consumption, access to solar technology
- Implementation and Maintenance: Technologies suited for installation and maintenance, both for network operators and end users
- Reliability: Simpler/Robust Equipment offering better reliability; Technologies less affected by weather or other environment conditions

Regulatory factors

In general, telecommunication regulations involve a complex domain of obligatory directives for a free and fair competitive telecom market within any country. An apex administrative body, like the "Bundesnetzagentur" in Germany, is usually responsible for enforcing such directives to protect the interests of the subscribers and prevent monopoly.

While regulatory issues involving frequency spectrum license fees/open access networks are fairly common; there are different regulatory concerns that need to be taken care of, for different technologies. Additional expenditure on account of regulatory requirements hinders the ambitions of network operators in rural broadband deployments. Fig. 2 shows the interaction between the parties involved and their concerns.

Digital Dividend: The radio frequency spectrum, (790-862 MHz) which has been derived out from the transition of terrestrial radio from analog to digital mode, is expected to be one of the important steps promoting Rural Broadband in Germany. Due to its physical wave propagation capabilities, this spectrum is particularly suited for supplying large areas with broadband utilizing lesser radio infrastructure.

Optical fibre based PON/AON: Open Access Network policies promote usage by a number of different providers that share the investments and maintenance cost. However, they come with an additional price of enhanced CAPEX which is around 20%-50% more than the actual CAPEX [6].

xDSL: In order to ensure a fair regulatory landscape, directives for ensuring unbundling of the local loop are enforced for network operators. This translates to an OPEX implication (multiple providers accessing the same copper loops). A suitable factor can be assumed for related calculations.

Socio-Economic factors

Economy status, revenue potential, actual demand for broadband services must be considered to construct the overall business case. Moreover, an account of funding/subsidy availability for the project must also be taken into account.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

C. *Rural Broadband – Global Developments*

Realizing the significance of broadband services for rural areas, different state aid programs and national strategies of most of the developing/developed economies have proposed substantial funding resources towards this cause:

Europe: The European Commission (EC) has approved ~ €1bn to be available for funding of rural broadband projects. The subsidy amount will be distributed among all Member States, which are responsible for project identification and documentation submission to EC for funding approval. Finally, the EC can fund up to 90% per project [7].The domains selected for funding involve: Creation of new broadband infrastructure including backhaul facilities (e.g., fixed, terrestrial wireless, satellite-based or combination of technologies); Upgrade of existing broadband infrastructure; Laying down passive broadband infrastructure (ducts, civil work, dark fibre). In particular for Germany, there is Broadband project, which aims to achieve 75% population coverage with at least 50 Mbit/s until 2014 [8]. Similarly for UK, the "Digital Britain" project proposes 2 Mbit/s for all users by the end of 2012.

USA: As part of the American Recovery and Reinvestment Act adopted in 2009, $7.2 billion was allocated for accelerating of deployment of broadband technologies in USA through the Broadband Stimulus Program. Two agencies, the Rural Utilities Service (RUS) under Department of Agriculture and National Telecommunications and Information Administration (NTIA) under Department of Commerce are responsible to distribute the money in terms of grants/loans to facilitate broadband deployment in rural areas and grants for deploying broadband infrastructure in un-served and underserved areas, enhance broadband capacity in public computer centers, thus, promote sustainable broadband adoption.

Asia-Pacific (APAC): The Federal Government of Australia has spent up to $258 million during its operation of connecting rural areas with broadband Internet services. The Department of Broadband, Communications and the Digital Economy established the program in 2007 as a subsidy to service providers for the setup cost of Internet connections that do not meet metro-comparable broadband speed benchmarks [9].

III. TECHNICAL OPTIONS

Although most of the developing/developed economies possess extensive copper cable infrastructure, with basic telephony services present almost in every nook and corner of the country; these copper based networks are limited in providing the requisite broadband services (through xDSL) due to longer distances between the DSLAM (Digital Subscriber Line Access Multiplexer) and the end-user. Moreover, rural and remote areas are generally characterized with no or limited telecommunication services, lower and limited economy, varying and rough geographical terrain, longer distances from COs (Central Office) of wireline networks or RBSs (Radio Base Station) of wireless networks. In such a scenario, a significantly large portion of the rural population is either un-served or underserved with respect to broadband Internet services.

Connecting rural areas with broadband Internet does not only involve evaluation of the potential technology candidates in terms of their technical capabilities but also considering the total value they bring along and the cost required to be paid in return. Thus, an extensive benchmarking of technical options for rural broadband is a necessary step before deciding on the potential choices for rural broadband solutions.

We adopt the following methodology to benchmark technical options based on a set of technical Key Performance Indicators (KPI), as follows:

- Maximum range
- Maximum throughput
- Next generation capabilities
- Quality of Service
- Interoperability
- Mobility
- Market status
- Innovation potential
- CAPEX/OPEX

While wireline networks are traditionally suited for high bandwidth data communications, wireless networks provide mobility support for voice and limited data requirements Technical evolutions in both these network classes promise almost comparable bandwidth intensive services with the possibility to be mobile. Ranking each technology on the basis of the aforementioned KPIs yielded the potential candidates for enabling broadband services in rural areas.

It is, however, important to note that in many cases, the maximum reach of a particular technology is already reached and serving a remote rural community requires a hybrid solution of technologies (for Backhaul & Access Network) to extend the maximum reach and bandwidth requirements.

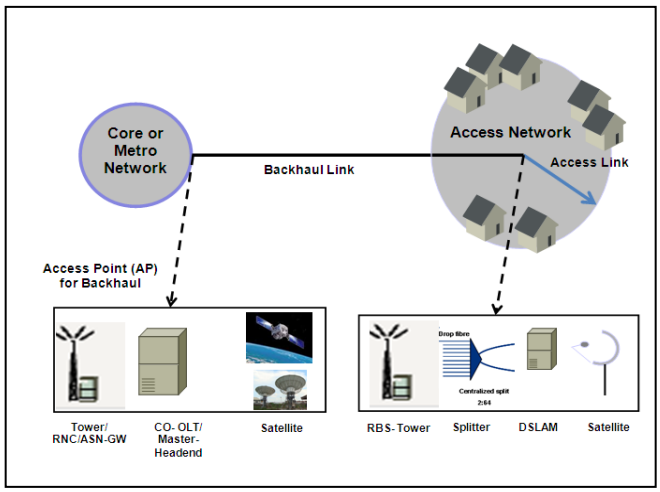


Figure 3. Hybrid Solutions for Rural Broadband- Backhaul+Access

Hence, the resulting combination solutions could be of the form- Optical Fibre (e.g. with PON) in the backhaul (connected to the CO which houses the OLT (Optical Line Terminal) at the Access Point for Backhaul Link) with WiMax as the access network technology or Microwave Point to Point as backhaul (connected to the RNC (Radio Network Controller) at the Access Point for Backhaul Link) and xDSL in the access region. These can be conveniently represented by the notation Optical Fibre+WiMax and Microwave P2P+VDSL respectively. We evaluate the cost-benefit implications for all of these resulting combinations, as illustrated in Fig. 3. This can then lead to an extended KPI set, particularly, for evaluating the hybrid technical solutions for rural broadband.

Technically, a number of possible alternatives can be utilized to connect rural communities. However, cost-wise, the choice of the scenario specific best-fitting solution is governed by factors like existing infrastructure, respective regulatory implications, revenue and funding opportunities. While abundant resources pointing towards the cost of network deployment [8, 9] are available, it is still difficult to comprehend the exact impact of these aforementioned factors affecting rural broadband costs and their interdependence.

IV. SOLUTION SELECTION STRATEGY

A. Process

A detailed analysis on any Rural Broadband deployment project yielded the following considerations that must be taken into account for complete cost-benefits modeling. The most significant heads along with their respective decisive parameters are as follows.

Technology feasibility analysis - Restricting the technical options to a set of feasible solutions for a given Capacity (Bandwidth) versus Distance (Reach) requirements (KPI based) for each technology.

Solution evaluation - Calculating costs (CAPEX/OPEX) for every feasible technology solution and **e**valuating them with respect to the following related aspects:

- *Existing infrastructure*: Deriving the value of the existing reusable resources
- *Regulatory inclusions and exemptions*: Establishing a cost towards regulatory obligations involved in the broadband technology deployment

Solution selection - Choice of the best-fitting technical solution in terms of costs, deployment feasibility and next generation network capabilities.

- *Revenue forecast*: Assessment of the market through a current estimate and forecasted revenue results
- *Funding resources*: Total Cost and State-dependent funding availability for the Rural Broadband project

A step-wise strategy, involving the aforementioned cost-related parameters that we developed to systematically model and evaluate the best solution, is illustrated in [1]. It would be worth noting that the following can be valid for a Greenfield or a Brownfield network deployment in rural areas. The value assigned for existing technical infrastructure in a Greenfield network is however, null.

B. Cost-Benefits Modeling

As is explicit, this step-wise approach comprehensively covers the most relevant aspects involving Rural Broadband and can also be utilized to obtain an accurate estimation of the Total Cost of Ownership of a project implementing any particular technology solution. This methodology laid the foundation for the development of the MS-Excel based Techno-Economic Model for Rural Broadband as part of this project work. It accepts the real scenario as the input and computes the Cost-Benefit values for the corresponding case as illustrated through Fig. 4.

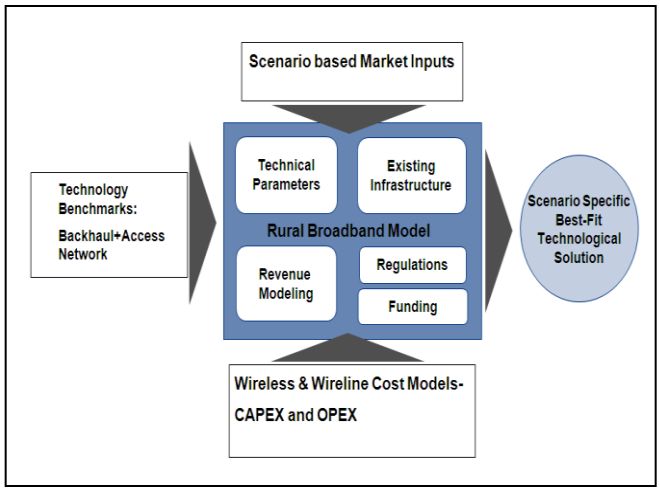


Figure 4. Cost Benefits Modeling for Rural Broadband Networks

C. Key Parameters & Objective Function

The cost-benefits value for any technical solution is modeled as NPC (Net project cost) expressed in € which encapsulates all the significant key parameters associated with rural broadband. NPC is essentially, the final cost required to be paid by a rural community in exchange for broadband services. The key parameters associated with modeling the cost-benefits of rural broadband are described as follows:

- CAPEX consists of the fixed infrastructure and equipment for a network operator or provider and it takes into account the number of targeted users, etc.
- OPEX deals with the costs for running the operations of a network operator during a certain period of time.
- Time period ("T") is the number of years, the project can be funded (Project duration)
- Regulatory cost implications ("REG")

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

- Existing infrastructure value deductions ("INF")
- Revenue per annum ("REV")
- Available funding amount for the rural broadband project ("FND")

Hence, the objective function describing the total costs payable by the rural community after including the relevant funding available, could be defined as NPC (Net Project Cost) which could have the following representation:

$$NPC(T) = CAPEX - INF + \sum_{\forall T} (OPEX + REG - REV) - FND \quad (1)$$

We adopt the aforementioned equation while implementing the Rural Broadband Cost-Benefits Model and it is worth mentioning that although REG represents a fixed cost value per annum; yearly values of OPEX and REV vary, depending on the rate of increase/decrease which could be defined in the quantitative model.

V. RESULTS AND ANALYSIS

We consider the example of a German community. It lies around 7 Km (Length of Backhaul Link) from the nearest district (Gemeinde) of Dietingen and is sparsely populated with just over 100 households (Access Network Demand). Although small, it is completely connected externally through the old POTS copper infrastructure, boasts of an accessible 2G Radio Base Station nearby and forms the prefect picture for any rural community which, although connected, is detached from the broadband network services. Gößlingen lies in the German state of Baden-Württemberg. Fig. 5 shows the funding potential in Germany and the rural case description.

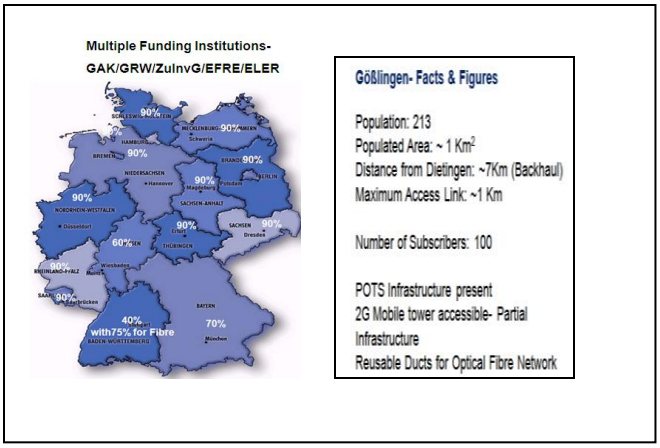


Figure 5. (a) Germany- Fundng Potential (b) Rural Case-Description

Although Germany has an extensive copper cable infrastructure, with basic telephony services present almost in every nook and corner of the country, these copper based networks are limited in providing the requisite broadband services (through xDSL) due to longer distances between the MDF (Main Distribution Frame)/ local exchange and the end- user. Thus, connecting any rural community today, not only involves developing the access network of the concerned region but also deploying or upgrading the link between the nearest access point (at the nearest town/district) and the community network, which we denote as "backhaul network" in this article.

Establishing or upgrading both these network segments, i.e. backhaul network and access network with technologies capable of providing broadband Internet at sufficient data rates to the end-user is of prime importance while working out the technical solution.

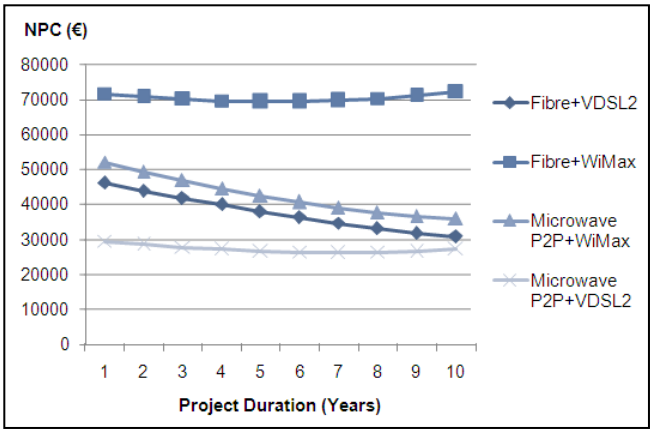


Figure 6. Comparing NPC for varied project duration

While optical fibre based wireline technologies have the potential to be sufficient for such scenarios, they come with an enormous price tag. Wireless technologies from the 3G/4G family and ones like WiMAX offer cheaper alternatives. However, longer stretches of the backhaul link (between the access point and the rural access network) and sufficiently higher access demands make these technologies unfavorable in many circumstances. Consequently, a combination of wireline and wireless technologies or mediums for the backhaul and access network seems to be a better proposition for this issue.

On plotting NPC with reference to the payback period (duration of the project) the following result, as illustrated in Fig.6, are typically obtained for the given example scenario. As explicit, the NPC is on a downward track for technical solutions which are profitable in the long run (less OPEX, more revenue). Also, utilizing the existing copper infrastructure through the VDSL2 access network is always advisable.

While results such as those obtained ín Fig. 6, must seem obvious for dense urban and suburban regions, it is important to note that they also hold good for rural networks with where demands as well as revenue potential are limited. The cost-revenue model developed during this project comprehensively considers all relevant parameters affecting the deployment costs of rural broadband network and can be

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

extended to other country scenarios with individual market, network, regulatory and funding data.

It should also be pointed out that the current project work was carried out to study the effect of all key parameters described in Fig.6. Moreover, this process based methodology can be extended to different settings and varied geographies.

VI. CONCLUSION

It is widely believed that providing broadband services for rural communities can be quite a challenging as well as an expensive assignment. However, solution models such as the one presented in this paper, can help presenting the total business case comprehensively and to reach a logical conclusion regarding the technical choice for network deployments. The key take aways include, but are not confined to:

- Reusing technical or civil infrastructure for Rural Broadband deployments can significantly reduce the costs of network establishment.
- A regulations friendly network implies larger investments. However, with sustained efforts on the part of the Regulator to promote Rural Broadband, some regulatory cost contributing aspects could be diminished or at least reduced by a substantial margin.
- In terms of a final technical solution for Rural Broadband; in addition to the cost of deployment, it is primarily important to consider the technical capabilities and NGN characteristics of the technologies before deciding on the potential candidates for the pool of technical solutions for Rural Broadband.

REFERENCES

[1] R. Zhao, N. Nayan, N. Zhelev, C. Mas Machuca, and W. Knospe, "Strategic Design for Rural Broadband Access Network", 5. ITG- Fachkonferenz, Breitband-versorgung in Deutschland, March 2011

[2] C. J. Kenny, "Expanding Internet Access to the rural poor in Africa", Information Technology for Development 9, 2000, pp. 25-31

[3] A. Kumar "Reduction in cost of rural network (CAPEX)" Course on Rural Telecommunications (ITU), Nov. 2005.

[4] Bundesministeriums für Wirtschaft und Technologie (BMWi): "The Federal Government's Broadband Strategy", [retrieved: 01/2012]

[5] R. L. Katz, S.Vaterlaus, P.Zenhäusern, S.Suter, and P.Mahler: " The impact of broadband on jobs and the German economy"

[6] Research Archive: Detecon International GmbH, Germany, http://www.detecon.com/, 2009-2010

[7] EUROPA: "EC MEMO/09/35", http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/09/35&format=HTML&aged=0&language=EN&guiLanguage=en [retrieved: 01/2012]

[8] Bundesministeriums für Wirtschaft und Technologie (BMWi): "Möglichkeiten der Breitbandförderung- Ein Leitfaden", www.bmwi.de, 2009

[9] Government of Australia eGovernment Resource Centre: http://www.egov.vic.gov.au/focus-on-countries/australia/trends-and-issues-australia/broadband-australia/audit-reveals-rural-broadband-subsidy-costs.html [retrieved: 01/2012]

[10] J. L. Riding, J. C. Ellershaw, A. V. Tran, L. J. Guan, and T. Smith, "Economics of Broadband Access Technologies for Rural Areas", Conference on Optical Fiber Communication, OFC 2009, pp. 1-3

[11] D.Žagar and V.Križanović, „Analyses and Comparisons of Technologies for Rural Broadband Implementation",17th International Conference on Software, Telecommunications & Computer Networks, SoftCOM 2009, pp. 292-296

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

On Fast Threefold Polarizations of Binary Discrete Memoryless Channels

Chengrong Huang, Ying Guo and Clement T. Gyamfi
School of Information Science & Engineering
Central South University
Changsha 410083, China
yingguo@csu.edu.cn

Tae Chul Shin and Moon Ho Lee
Institute of Information and Communication
Chonbuk National University
Chonju 561-756, Korea
moonho@chonbuk.ac.kr

***Abstract*—Motivated by a polarization approach to construct code sequences related to Reed-Muller (RM) codes with generator matrix G_{2^n} of size $2^n \times 2^n$ to increase the cutoff rate, we consider a problem of systematic constructions of polar codes as splitting threefold input binary discrete memoryless channels (B-DMC) for generator matrix G_{3^n}. The polarized channel achieves the symmetric capacity of arbitrary binary-input discrete memoryless channels under a low computation complexity of successive cancellation decoding strategy for any core matrix $\mathcal{O}_3$, which is a submatrix of generator matrix $G_4 = \mathcal{O}_2 \otimes \mathcal{O}_2$. In principle larger matrices G_{3^n} with fast construction algorithms can be used for constructions of polar code sequences that tend to polarize with respect to the rate and reliability under certain fast combining and splitting operations. The proposed polarization code scheme has a salient recursiveness feature and hence can be decoded with a belief propagation (BP) decoder, which renders the scheme analytically tractable and provides a powerful low-complexity coding algorithm.**

Keywords-polar codes; binary discrete memoryless channels; channel coding; fast algorithm.

I. INTRODUCTION

The channel polarization may be consisted of code sequences using a belief propagation (BP) decoder with symmetric high rate capacity in given binary-input discrete memoryless channels (B-DMC) [1]. It is a commonplace phenomenon that is almost impossible to avoid as long as several similar channels are combined in a sufficient density with certain elegant connections. The investigation of channel polarization not only has become an interesting theoretical problem, but also have lots of practical applications in signal sequence transforms, data processing, signal processing, and code coding theory [2], [3].

Motivated by a fascinating aspect of Shannon's channel coding theorem that shows the existence of capacity-achieving code sequences [4], we show a novel construction of provably capacity-achieving sequences with low coding complexities with BP decoders. This paper is an attempt to meet this elusive goal for B-DMC, which is an extension of work where channel combining and splitting were used to to improve the sum cutoff rate [1]–[3]. In a recent investigation, the above-mentioned rate has been generalized for different forms of polar-code constructions [5]. However, there is few recursive method suggested there to reach the ultimate limit of such improvements. As the present work progressed, it is shown that polar-code sequences have much in common with Reed-Muller codes [6]. Indeed, recursive code construction and successive cancellation decoding, which are two essential characters of polar coding, appear to be introduced into coding theory. It has a relationship to existing work by noting that polar-code sequences can be made to be multilevel in terms of generator matrices G_{p^n} originating from Plotkin's constructions [7]. Therefore, Polar coding has a strong resemblance to Reed-Muller coding, and hence may be regarded as a generalization of Reed-Muller codes since both coding constructions start with a generator matrix for a rate one code and obtain generator matrices of lower rate codes by expurgating rows of the initial generator matrices. While in this paper, we would like to point out that polar-code sequences that have the same structure as Reed-Muller codes have a sparse factor graph representation and can be fast decoded with BP decoder for superior performance [3], [8].

Since polar-coding, which may be considered as a generalization of Reed-Muller coding, is an approach employed to construct capacity-achieving codes with certain symmetries, we demonstrate the performance advantages of several polar-code sequences under BP decoder with respect to symmetric capacity and Bhattacharyya parameter. The symmetric capacity is the highest rate achievable subject to using the input alphabets of B-DMC with equal probability. Polar-code is the first provably capacity achieving code with low coding complexity [9].

According to construction of polar-code sequences, we consider a generic B-DMC denoted by $W : \mathcal{X} \mapsto \mathcal{Y}$ with input alphabets $\mathcal{X} = \{0, 1\}$, output alphabets $\mathcal{Y}$, and transition probabilities $W(y|x)$ for $x \in \mathcal{X}, y \in \mathcal{Y}$. There are two channel parameters [1], i.e., the symmetric capacity

$$I(W) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \frac{1}{2} W(y|x) \log \frac{W(y|x)}{\frac{1}{2}W(y|0) + \frac{1}{2}W(y|1)}. \quad (1)$$

and the Bhattacharyya parameter

$$Z(W) = \sum_{y \in \mathcal{Y}} \sqrt{W(y|0)W(y|1)}. \quad (2)$$

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

The two parameters are much useful as measures of rate and reliability of the B-DMC, i.e., the Shannon capacity $I(W)$ is the highest rate at which reliable communication is possible using the inputs with equal frequency, and $Z(W)$ is an upper bound on the probability of maximum-likelihood (ML) decision error.

Throughout this paper, we use the notation $\mathbf{a}_1^N$ to denote a row vector $(a_1, \cdots, a_N)$. Given such a vector $\mathbf{a}_1^N$, we write $\mathbf{a}_i^j$ to denote the subvector $(a_i, \cdots, a_j)$. Moreover, we write $\mathbf{a}_{\mathcal{A}}$ to denote the subvector $(a_i : i \in \mathcal{A} \subseteq \{1, 2, \cdots, N\})$. We write $\mathbf{a}_{1,o}^j$ to denote the subvector with odd indices $(a_i : 1 \leq i \leq j,\ i \text{ odd })$, and $\mathbf{a}_{1,e}^j$ to denote the subvector with even indices $(a_i : 1 \leq i \leq j,\ i \text{ even })$. Similarly, we write $\mathbf{a}_{1,l}^j$ to denote the subvector with the indices $(a_i : 1 \leq i \leq j,\ i = pk + l)$. We write W^N to denote the channel corresponding to N uses of B-DMC W, and hence, $W^N : \mathcal{X}^N \mapsto \mathcal{Y}^N$ with $W^N(\mathbf{y}_1^N|\mathbf{x}_1^N) = \prod_{i=1}^N W(y_i|x_i)$.

This paper is organized as follows. In Sec.II, generation matrices of polar codes are presented via the channel combining and splitting. In Sec.III, according to the properties of polarization constructions, a decoding algorithm is suggested for the G_N-coset codes. Finally, conclusions are drawn in Sec.III.

II. Polarization Construction

In this section, we derive fast constructions of polar-codes based on Arikan's construction [1]. We begin by giving an explicit algebraic expression of generator matrix G_N of polar-code, which has been defined in a schematic form. The algebraic form of G_N point at an efficient implementation of coding operation $\mathbf{u}_1^N G_N$. In analyzing the coding operation, we exploit its relation to fast transforms in signal processing [9].

We carry out the construction of G_N-coset codes before specializing polar-codes. Recall that individual G_N-coset codes are identified by a parameter vector $(N, K, \mathcal{A}, \mu_{\mathcal{A}^c})$ [1]. In the following analysis, we fix the shorted parameter vector $(N, K, \mathcal{A})$ while keeping free $\mu_{\mathcal{A}^c}$ to take any value over $\mathcal{X}^{N-K}$ as frozen bits. In other words, the analysis of polar-code sequences will be over the ensemble of G_N-coset codes with a fixed parameter vector $(N, K, \mathcal{A})$ based on several families of generator matrices $G_N = \mathcal{O}_3^{\otimes n}$, where $\otimes$ denotes keronecker product, n is a positive integer.

Constructions of polar-code sequences based on generator matrices G_{3^n} are derived from the radix $N = 3^n$ channel polarization, which is an operation by which one manufacture out of N independent copies of a given B-DMC W yields a second set of N channels $\{W_N^i : 1 \leq i \leq N\}$ that show a polarization effect in a sense that, as N becomes large, the symmetric capacity terms $\{I(W_N^i) : 1 \leq i \leq N\}$ tend towards 0 or 1 for all but a vanishing fraction of indices i. This operation consists of a channel combining phase and a channel splitting phase.

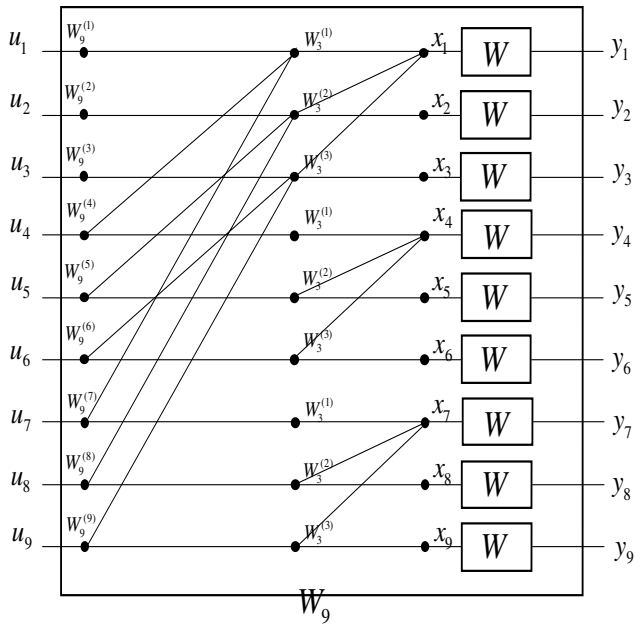


Figure 1. Transformation of $\mathcal{O}_9 = (\mathcal{O}_3 \otimes I_3)(I_3 \otimes \mathcal{O}_3)$.

Taking block length $N = 3^n$, the channel combining based on core matrix $\mathcal{O}_3$ of order 3 includes 3^n copies of a given B-DMC W in a recursive manner to produce a vector channel $W_{3^n}^{(i)}$ for any $1 \leq i \leq 3^n$. In a similar way , the first level of the recursion combines three independent copies of W as shown in Fig. 1 and achieves the combined channel W_3 with the transition probabilities described as

$$W_3(\mathbf{y}_1^3|\mathbf{u}_1^3) = W(y_1| \oplus_{i=1}^3 u_i)W(y_2|u_2)W(y_3|u_3), \quad (3)$$

where the mapping W_3 is defined as $W_3(\mathbf{y}_1^3|\mathbf{x}_1^3) = W^3(\mathbf{y}_1^3|\mathbf{u}_1^3\mathcal{O}_3)$, where the core matrix

$$\mathcal{O}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix},$$

is sub-matrix of $\mathcal{O}_4 = \mathcal{O}_2^{\otimes 2}$, i.e.,

$$\mathcal{O}_4 = \begin{pmatrix} \mathcal{O}_3 & \mathbf{0}_{3\times 1} \\ \mathbf{1}_{1\times 3} & 1 \end{pmatrix},$$

where $\mathbf{1}_{1\times 3} = (1, 1, 1)$ and $\mathbf{0}_{3\times 1} = (0, 0, 0)^T$.

For the second level of the recursion, we combine three independent copies of W_3, as shown in Fig. 2, to generate the channel W_{3^2} with transition probabilities

$$W_{3^2}(\mathbf{y}_1^9|\mathbf{u}_1^9) = W_3(y_1^3| \oplus_{i=1}^3 u_i, \oplus_{i=4}^6 u_i, \oplus_{i=7}^9 u_i) \cdot W_3(y_4^6|u_2, u_5, u_8)W_3(y_7^9|u_3, u_6, u_9). \quad (4)$$

Define the permutation operation $B_9 = R_9$, i.e.,

$$B_9(\mathbf{u}_1^9) = (u_1, u_4, u_7, u_2, u_5, u_8, u_3, u_6, u_9). \quad (5)$$

Therefore, we obtains the mapping $\mathbf{u}_1^9 \rightarrow \mathbf{x}_1^9$ from the input of W_9 to the input of W^9 such that $\mathbf{x}_1^9 = \mathbf{u}_1^9 G_9$, where $G_9 = B_9\mathcal{G}_9 = B_9\mathcal{O}_9 = B_9\mathcal{O}_3^{\otimes 2}$.

Generally, we get the extensive form of the recursion while three independent copies of $W_{3^{n-1}}$ are combined to produce the channel W_{3^n}. The input vector $\mathbf{u}_1^{3^n}$ is transformed to $\mathbf{s}_1^{3^n}$ such that

$$s_{3i-2} = \oplus_{j=0}^{2} u_{3i-j}, \quad s_{3i-1} = u_{3i-1}, \quad s_{3i} = u_{3i}$$

for $1 \leq i \leq 3^{n-1}$. The operator R_{3^n} is a permutation operation defined as

$$\begin{aligned} R_{3^n}(\mathbf{u}_1^{3^n}) &= (\mathbf{u}_{1,1}^{3^n}, \mathbf{u}_{1,2}^{3^n}, \mathbf{u}_{1,3}^{3^n}) \\ &= (u_1, \cdots, u_{3^n-2}, u_2, \cdots, u_{3^n-1}, u_3, \cdots, u_{3^n}). \end{aligned}$$

It is obvious that the mapping $\mathbf{u}_1^{3^n} \rightarrow \mathbf{x}_1^{3^n}$ from the input of the synthesized channel W_{3^n} to the input of the underlying raw channels W^{3^n} is linear and hence can be represented with a generator matrix G_{3^n} so that $\mathbf{x}_1^{3^n} = \mathbf{u}_1^{3^n} G_{3^n}$. Thus the relation of transition probabilities of W_{3^n} and W^{3^n} are described as $W_{3^n}(\mathbf{y}_1^{3^n}|\mathbf{u}_1^{3^n}) = W^{3^n}(\mathbf{y}_1^{3^n}|\mathbf{u}_1^{3^n} G_{3^n})$, where $\mathbf{y}_1^{3^n} \in \mathcal{Y}^{3^n}$, $\mathbf{u}_1^{3^n} \in \mathcal{X}^{3^n}$, $G_{3^n} = B_{3^n} G_3^{\otimes n}$ and B_{3^n} is a 3^n-order permutation matrix defined by $B_{3^n} = R_{3^n}(I_3 \otimes B_{3^{n-1}})$.

According to the previously defined processing for channel combining and splitting which transforms 3 independent copies of W into $W_3^{(i)}$ for $1 \leq i \leq 3$, we get the following one-to-one mapping to describe the relation of W and $W_3^{(i)}$ $\Xi_3 : (W, W, W) \mapsto (W_3^{(1)}, W_3^{(2)}, W_3^{(3)})$, where

$$W_3^{(1)}(\mathbf{y}_1^3|u_1) = \sum_{u_2,u_3} \frac{1}{3} W(y_1| \oplus_{i=1}^{3} u_i) W(y_2|u_2) W(y_3|u_3)$$

$$W_3^{(2)}(\mathbf{y}_1^3, u_1|u_2) = \sum_{u_3} \frac{1}{3} W(y_1| \oplus_{i=1}^{3} u_i) W(y_2|u_2) W(y_3|u_3)$$

$$W_3^{(2)}(\mathbf{y}_1^3, \mathbf{u}_1^2|u_3) = \frac{1}{3} W(y_1| \oplus_{i=1}^{3} u_i) W(y_2|u_2) W(y_3|u_3).$$

In a similar way, for $N = 3^n$ we achieve the generalized mapping to establish the relation of $W_N^{(i)}$ and $W_{3N}^{(k)}$ as follows

$$\Xi_{3^i} : (W_N^{(i)}, W_N^{(i)}, W_N^{(i)}) \mapsto (W_{3N}^{(3i-2)}, W_{3N}^{(3i-1)}, W_{3N}^{(3i)}),$$

where

$$\begin{aligned} &W_{3N}^{(3i-2)}(\mathbf{y}_1^{3N}, \mathbf{u}_1^{3i-3}|u_{3i-2}) \\ &= \sum_{u_{3i-1}, u_{3i}} \frac{1}{3} W_N^{(i)}(\mathbf{y}_1^N, \oplus_{j=1}^{3} \mathbf{u}_{1,j}^{3i-3} | \oplus_{j=0}^{2} u_{3i-j}) \\ &\cdot W_N^{(i)}(\mathbf{y}_{N+1}^{2N}, \mathbf{u}_{1,2}^{3i-3}|u_{3i-1}) W_N^{(i)}(\mathbf{y}_{2N+1}^{3N}, \mathbf{u}_{1,3}^{3i-3}|u_{3i}) \end{aligned}$$

$$\begin{aligned} &W_{3N}^{(3i-1)}(\mathbf{y}_1^{3N}, \mathbf{u}_1^{3i-2}|u_{3i-1}) \\ &= \sum_{u_{3i}} \frac{1}{3} W_N^{(i)}(\mathbf{y}_1^N, \oplus_{j=1}^{3} \mathbf{u}_{1,j}^{3i-3} | \oplus_{j=0}^{2} u_{3i-j}) \\ &\cdot W_N^{(i)}(\mathbf{y}_{N+1}^{2N}, \mathbf{u}_{1,2}^{3i-3}|u_{3i-1}) W_N^{(i)}(\mathbf{y}_{2N+1}^{3N}, \mathbf{u}_{1,3}^{3i-3}|u_{3i}) \end{aligned}$$

$$\begin{aligned} &W_{3N}^{(3i)}(\mathbf{y}_1^{3N}, \mathbf{u}_1^{3i-1}|u_{3i}) \\ &= \frac{1}{3} W_N^{(i)}(\mathbf{y}_1^N, \oplus_{j=1}^{3} \mathbf{u}_{1,j}^{3i-3} | \oplus_{j=0}^{2} u_{3i-j}) \\ &\cdot W_N^{(i)}(\mathbf{y}_{N+1}^{2N}, \mathbf{u}_{1,2}^{3i-3}|u_{3i-1}) W_N^{(i)}(\mathbf{y}_{2N+1}^{3N}, \mathbf{u}_{1,3}^{3i-3}|u_{3i}). \end{aligned}$$

The transformation Ξ_{3^n} is not only rate-preserving but also reliability-improving, i.e.,

$$\sum_{j=0}^{2} I(W_{3N}^{(3i-j)}) = 3I(W_N^{(i)})$$

$$\sum_{j=0}^{2} Z(W_{3N}^{(3i-j)}) \leq 3Z(W_N^{(i)}).$$

In addition, the channel splitting moves the rate and reliability away from the center. Namely, we obtain the following results

$$\begin{aligned} &I(W_{3N}^{(3i-2)}) \leq I(W_{2N}^{(3i-1)}) \leq I(W_N^{(i)}) \leq I(W_{3N}^{(3i)}) \\ &Z(W_{3N}^{(3i-2)}) \geq Z(W_{3N}^{(3i-1)}) \geq Z(W_N^{(i)}) \geq Z(W_{3N}^{(3i)}). \end{aligned}$$

The afore-mentioned reliability terms further satisfy the following constrained conditions

$$\begin{aligned} &Z(W_{3N}^{(3i-2)}) \leq 3Z(W_N^{(i)}) - 2Z^3(W_N^{(i)}) \\ &Z(W_{3N}^{(3i-1)}) = Z(W_{3N}^{(3i)}) = Z(W_N^{(i)})^3. \end{aligned} \quad (6)$$

To illustrate the process of polarization on the basis of core matrix $\mathcal{O}_3$ for a given $N = 3^n$, each input sequence $\mathbf{u}_1^N$ can be encoded through using an encoder

$$\mathbf{x}_1^N = \mathbf{u}_1^N G_N, \quad (7)$$

where $G_N = B_N \mathcal{O}_3^{\otimes n}$ is the generator matrix of order N and B_N is a permutation matrix (operation) defined in the recursion way as

$$B_N = R_N(I_3 \otimes R_{N/3}) \cdots (I_{N/3} \otimes R_3). \quad (8)$$

We note that $\mathcal{O}_3 \cdot \mathcal{O}_3 = I_3$ and $B_N = R_N(I_3 \otimes B_{N/3})$.

Actually, it is easy to prove that $R_N(G_3 \otimes I_{N/3}) = (I_{N/3} \otimes \mathcal{O}_3)R_N$. Therefore, one has

$$G_N = (I_{N/3} \otimes G_3)R_N(I_3 \otimes R_{N/3}),$$

which can be rewritten as

$$\begin{aligned} G_N &= R_N(\mathcal{O}_3 \otimes G_{N/3}) \\ &= R_N(I_3 \otimes R_{N/3})(\mathcal{O}_3^{\otimes 2} \otimes G_{N/3^2}). \end{aligned} \quad (9)$$

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Denote $\mathcal{G}_N = \mathcal{O}_3^{\otimes n}$. Then one achieves

$$\begin{aligned}\mathcal{G}_N &= \mathcal{G}_{N/3} \otimes \mathcal{G}_3 \\ &= \prod_{i=1}^{n}(I_{3^{n-i}} \otimes \mathcal{O}_3 \otimes I_{3^{i-1}}) = \prod_{i=1}^{n}\mathcal{G}_N^i, \quad (10)\end{aligned}$$

where $\mathcal{G}_N^i = I_{3^{n-i}} \otimes \mathcal{G}_3 \otimes I_{3^{i-1}}$ and $G_3 = \mathcal{O}_3$. Consequently there are N row's permutation matrices $P_N^r(i)$ and N column's permutation matrices $P_N^c(i)$ of $\mathcal{G}_N^i$ such that $P_N^r(i) \cdot P_N^c(i) = P_N^c(i) \cdot P_N^r(i)$. Then we show that the factorizations have equal factors as follows

$$P_N^r G_N P_N^c = \prod_{i=1}^{n}\hat{\mathcal{G}}_N^i = (I_{3^{n-1}} \otimes \mathcal{O}_3)^n, \quad (11)$$

where $\hat{\mathcal{G}}_N^i = P_N^r(i) \cdot \mathcal{G}_N^i \cdot P_N^r(i)$.

Moreover, we consider another channel combining scheme based on core matrix

$$\hat{\mathcal{O}}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

to generate a vector channel $W_{3^n}^{(i)}$, where the core matrix $\hat{\mathcal{O}}_3$ is sub-matrix of $\mathcal{O}_4 = \mathcal{O}_2^{\otimes 2}$, i.e.,

$$\mathcal{O}_4 = \begin{pmatrix} 1 & \mathbf{0}_{3\times 1} \\ \mathbf{1}_{1\times 3} & \hat{\mathcal{O}}_3 \end{pmatrix},$$

where $\mathbf{1}_{1\times 3} = (1,1,1)$, $\mathbf{0}_{3\times 1} = (0,0,0)^T$, and $\hat{\mathcal{O}}_3 \cdot \hat{\mathcal{O}}_3 = I_3$. In this case the first level of the recursion combines three independent copies of W that achieves the combined channel W_3 with the following transition probabilities

$$W_3(\mathbf{y}_1^3|\mathbf{u}_1^3) = W(y_1|u_1 \oplus u_3)W(y_2|u_2 \oplus u_3)W(y_3|u_3), \quad (12)$$

where W_3 is defined as $W_3(\mathbf{y}_1^3|\mathbf{x}_1^3) = W^3(\mathbf{y}_1^3|\mathbf{u}_1^3\hat{\mathcal{O}}_3)$. To design the second level of the recursion we obtain the combined channel W_{3^2} in Fig. 2 with transition probabilities

$$\begin{aligned}W_{3^2}(\mathbf{y}_1^9|\mathbf{u}_1^9) &= W_3(y_1^3|u_1 \oplus u_3, u_4 \oplus u_6, u_7 \oplus u_9) \\ &\cdot W_3(y_4^6| \oplus_{i=2}^3 u_i, \oplus_{i=5}^6 u_i, \oplus_{i=8}^9 u_i)W_3(y_7^9|u_3, u_6, u_9).\end{aligned}$$

To construct polar-code sequences of block length 3^n based on the polarization of channel with generator matrices G_{3^n} for core matrix $\mathcal{O}_3$ of order 3, we should compute the reliability channel polarization in terms of the vector

$$Z(3^n) = (Z_{3^n,1}, Z_{3^n,2}, \cdots, Z_{3^n,3^n})$$

through using the recursion

$$Z_{3k,j} = \begin{cases} 3Z_{k,j} - 2Z_{k,j}^3, & \text{for } 1 \le j \le k; \\ Z_{k,j-k}^3, & \text{for } k+1 \le j \le 2k, \\ Z_{k,j-2k}^3, & \text{for } 2k+1 \le j \le 3k, \end{cases} \quad (13)$$

for any $k = 1, 3, 3^2, \cdots, 3^{n-1}$ starting with $Z_{1,1} = 1/2$. After that we generate a permutation operation $\pi_{3^n} = (i_1, \cdots, i_{3^n})$ with respect to the set $(1, \cdots, 3^n)$ so that $Z_{3^n,i_j} < Z_{3^n,i_k}$ for any $1 \le j < k \le 3^n$. The generator matrix $\mathcal{G}_P(3^n, K)$ of a $(3^n, K)$ polar-code can be constructed from the sub-matrix of G_{3^n} with indices $\{i_1, \cdots, i_K\} \subseteq \{1, \cdots, N\}$. According to the polarization of channel with generator matrices G_{3^n}, it is obvious that the computational complexity of this processing is $2n3^{\frac{n}{2}}$. However, the computational complexity of the direct approach is $n3^n$, which shows an advantage of the proposed construction.

Example 1: Taking the matrix $\mathcal{G}_9 = \mathcal{O}_3^{\otimes 2}$ for recursion in Eq.(13), we have

$$\begin{aligned}Z_9 = (&0.857, 0.034, 0.034, 0.034, 0.001, 0.001, \\ &0.034, 0.001, 0.001), \quad (14)\end{aligned}$$

which gives the permutation $\pi_8 = (9,8,6,5,7,4,3,2,1)$ for rows of the generator matrix $\mathcal{G}_9$. Exploiting the polarization of channel with generator matrices G_9, the code $(9, 5, \{9,8,6,5,7\}, (0,0,0,0)$ can be constructed with the encoder mapping as follows

$$\begin{aligned}\mathbf{x}_1^9 &= \mathbf{u}_1^9 G_9 \\ &= (u_{98657}) \begin{pmatrix} 110110000 \\ 101101000 \\ 100100100 \\ 110110110 \\ 101010101 \end{pmatrix} + (0000) \begin{pmatrix} 100000000 \\ 110000000 \\ 101000000 \\ 000100000 \end{pmatrix}.\end{aligned}$$

For a source block $(1,1,1,1)$, the coded block is $\mathbf{x}_1^9 = (1,1,1,1,1,0,0,0,0)$. It is necessary to note that this code is an $(N, K) = (9, 5)$ Reed-Muller code with the generator matrix

$$G_R = \begin{pmatrix} 110110000 \\ 101101000 \\ 100100100 \\ 110110110 \\ 101010101 \end{pmatrix}.$$

III. Decoding Algorithm

In this section, we consider the decoding algorithm of the proposed polar codes. As in the previous section, our computational model will be a single processor machine with a random-access memory. We consider the decoding of G_N-coset codes with parameters $(N, K, \mathcal{A}, \mu_{\mathcal{A}^c})$ for a given block length $N = 3^n$.

Recall that the source vector $\mathbf{u}_1^N$ consists of a random part $\mu_{\mathcal{A}}$ and a frozen part $\mu_{\mathcal{A}^c}$ such that $\mathbf{u}_1^N = \{\mu_{\mathcal{A}} \cup \mu_{\mathcal{A}^c}\}$. This vector $\mathbf{u}_1^N$ is transmitted across W_N and a channel output $\mathbf{y}_1^N$ is obtained with probability $W_N(\mathbf{y}_1^N|\mathbf{u}_1^N)$. The decoder observes $(\mathbf{y}_1^N, \mu_{\mathcal{A}^c})$ and generates an estimate $\hat{\mathbf{u}}_1^N$ of $\mathbf{u}_1^N$.

If $i \in \mathcal{A}^c$, the element u_i is known, and thus the i-th decision element is $\hat{u}_i = u_i$. However, if $i \in \mathcal{A}$, then the i-th decision element waits until it has received the previous decisions $\hat{\mathbf{u}}_1^{i-1}$. Upon receiving them, the decoder computes

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

$$
\begin{aligned}
L_N^{(3i-2)}(\mathbf{y}_1^N, \hat{\mathbf{u}}_1^{3i-3}) &= \frac{W_N(\mathbf{y}_1^N, \hat{\mathbf{u}}_1^{3i-3}|0)}{W_N(\mathbf{y}_1^N, \hat{\mathbf{u}}_1^{3i-3}|1)} \\
&= \frac{L_{N/3}^{(i)}(\mathbf{y}_1^{N/3}, \oplus_{j=1}^3 \hat{\mathbf{u}}_{1,j}^{3i-3}) L_{N/3}^{(i)}(\mathbf{y}_{2N/3+1}^N, \hat{\mathbf{u}}_{1,3}^{3i-3}) [L_{N/3}^{(i)}(\mathbf{y}_{N/3+1}^{2N/3}, \hat{\mathbf{u}}_{1,2}^{3i-3}) + 1]}{L_{N/3}^{(i)}(\mathbf{y}_1^{N/3}, \oplus_{j=1}^3 \hat{\mathbf{u}}_{1,j}^{3i-3}) + L_{N/3}^{(i)}(\mathbf{y}_{N/3+1}^{2N/3}, \hat{\mathbf{u}}_{1,2}^{3i-3}) + L_{N/3}^{(i)}(\mathbf{y}_{2N/3+1}^N, \hat{\mathbf{u}}_{1,3}^{3i-3}) + 1} \\
&\quad + \frac{L_{N/3}^{(i)}(\mathbf{y}_{N/3+1}^{2N/3}, \hat{\mathbf{u}}_{1,2}^{3i-3}) [L_{N/3}^{(i)}(\mathbf{y}_1^{N/3}, \oplus_{j=1}^3 \hat{\mathbf{u}}_{1,j}^{3i-3}) + L_{N/3}^{(i)}(\mathbf{y}_{2N/3+1}^N, \hat{\mathbf{u}}_{1,3}^{3i-3})]}{L_{N/3}^{(i)}(\mathbf{y}_1^{N/3}, \oplus_{j=1}^3 \hat{\mathbf{u}}_{1,j}^{3i-3}) + L_{N/3}^{(i)}(\mathbf{y}_{N/3+1}^{2N/3}, \hat{\mathbf{u}}_{1,2}^{3i-3}) + L_{N/3}^{(i)}(\mathbf{y}_{2N/3+1}^N, \hat{\mathbf{u}}_{1,3}^{3i-3}) + 1} \\
L_N^{(3i-1)}(\mathbf{y}_1^N, \hat{\mathbf{u}}_1^{3i-2}) &= \frac{W_N(\mathbf{y}_1^N, \hat{\mathbf{u}}_1^{3i-2}|0)}{W_N(\mathbf{y}_1^N, \hat{\mathbf{u}}_1^{3i-2}|1)} \\
&= \frac{L_{N/3}^{(i)}(\mathbf{y}_{N/3+1}^{2N/3}, \hat{\mathbf{u}}_{1,2}^{3i-3}) [L_{N/3}^{(i)}(\mathbf{y}_1^{N/3}, \oplus_{j=1}^3 \hat{\mathbf{u}}_{1,j}^{3i-3})^{1-2\hat{u}_{3i-2}} L_{N/3}^{(i)}(\mathbf{y}_{2N/3+1}^N, \hat{\mathbf{u}}_{1,3}^{3i-3}) + 1]}{L_{N/3}^{(i)}(\mathbf{y}_{2N/3+1}^N, \hat{\mathbf{u}}_{1,3}^{3i-3}) + L_{N/3}^{(i)}(\mathbf{y}_1^{N/3}, \oplus_{j=1}^3 \hat{\mathbf{u}}_{1,j}^{3i-3})^{1-2\hat{u}_{3i-2}}} \\
L_N^{(3i)}(\mathbf{y}_1^N, \hat{\mathbf{u}}_1^{3i-1}) &= \frac{W_N(\mathbf{y}_1^N, \hat{\mathbf{u}}_1^{3i-1}|0)}{W_N(\mathbf{y}_1^N, \hat{\mathbf{u}}_1^{3i-1}|1)} \\
&= L_{N/3}^{(i)}(\mathbf{y}_1^{N/3}, \oplus_{j=1}^3 \hat{\mathbf{u}}_{1,j}^{3i-3})^{1-2(\hat{u}_{3i-2} \oplus \hat{u}_{3i-1})} L_{N/3}^{(i)}(\mathbf{y}_{N/3+1}^{2N/3}, \hat{\mathbf{u}}_{1,2}^{3i-3})^{1-2\hat{u}_{3i-1}} L_{N/3}^{(i)}(\mathbf{y}_{2N/3+1}^N, \hat{\mathbf{u}}_{1,3}^{3i-3})
\end{aligned} \tag{15}
$$

the likelihood ratio (LR) as follows

$$
L_N^{(i)}(\mathbf{y}_1^N, \hat{\mathbf{u}}_1^{i-1}) = \frac{W(\mathbf{y}_1^N, \hat{\mathbf{u}}_1^{i-1}|0)}{W(\mathbf{y}_1^N, \hat{\mathbf{u}}_1^{i-1}|1)}, \tag{16}
$$

and generates its decision through using

$$
\hat{u}_i = \begin{cases} 0, & \text{if } L_N^{(i)}(\mathbf{y}_1^N, \hat{\mathbf{u}}_1^{i-1}) \geq 1; \\ 1, & \text{otherwise,} \end{cases} \tag{17}
$$

which is then sent to all succeeding decision elements. This processing is a single-pass algorithm, with no revision of estimates. The complexity of this algorithm is determined essentially by the complexity of computing the LRs.

As for polarization of channel based on generator matrix G_{3^n} of order 3^n, we calculate with the recursive formulas (6)-(6) on the top of the next page based on core matrix $\mathcal{O}_3$ and obtain the formula in Eq.(15).

IV. Conclusion

In this paper, we considered the overall encoding/decoding structures and systems of the polar-code sequence to show the expression of encoding/decoding of the polar-code with fast algorithms based on its generator matrix $\mathcal{G}_{3^n}$. The complexity of the proposed encoding scheme is much lower than the previous which is proposed by Arikan. By transmitting the information bits over the almost noiseless B-DMC W, polar-codes of block-length 3^n can be fast constructed starting with any polarizing matrix G_{3^n}. The encoding and successive cancellation decoding complexities of such codes are lower than Arikan's code.

Acknowledgment

This work was supported by World Class University R32-2010-000-20014-0 NRF, FR2010-0020942 NRF, Post BK21, Korea, 2011 Korea-China international Cooperative Research Project(Grant Nos, Dooo66,100026), and partly by NSFC 60902044, the New Century Excellent Talents in University (NCET-11-0510), China.

References

[1] E. Arikan, Channel Polarization: A Method for Constructing Capacity-Achieving codes for Symmetric Binary-Input Memoryless Channel, *IEEE Trans. Inform. Theory*, Vol. 55, No. 7, July 2009.

[2] E. Arikan, Channel Combining and Splitting for Cutoff Rate Improvement, *IEEE Trans. Inform. Theory*, Vol. 52, No. 2, FEB 2006.

[3] E. Arikan, performance comparison of polar-codes and Reed-Muller codes, *IEEE. Comm. Lett.*, Vol. 12, pp. 447-449, June 2008.

[4] C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.*, Vol. 27, pp. 379-423, 623-656, Jul.-Oct. 1948.

[5] E. Arikan and E. Telatar, the rate of channel polarization, July 2008. [Online]. Available: arXiv:0807.3806v2.

[6] I. Reed, A class of multiple-error-correcting codes and the decoding scheme, *IRE Trans. Inf. Theory*, Vol. 4, No. 3, pp. 39-44, Sep. 1954.

[7] M. Plotkin, Binary codes with specified minimum distance, *IRE Trans. Inf. Theory*, Vol. 6, No. 3, pp. 445-450, Sep. 1960.

[8] G. D. Forney Jr., Codes on graphs: Normal realizations, *IEEE Trans. Inform. Theory*, Vol. 47, No. 2, pp. 520-548, Feb. 2001.

[9] G. Zeng and M. H. Lee, A Generalized Reverse Block Jacket Transform, *IEEE Trans. Circuits Syst. I*, Vol. 55, No. 6, pp. 651-669, 2008.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Optimal Allocation of Fibre Delay Lines in Optical Burst Switched Networks

Daniele Tafani
The Rince Institute
School of Electronic Engineering
Dublin City University
Glasnevin, Dublin 9
Email: tafanid@eeng.dcu.ie

Conor McArdle
The Rince Institute
School of Electronic Engineering
Dublin City University
Glasnevin, Dublin 9
Email: mcardlec@eeng.dcu.ie

Liam P. Barry
The Rince Institute
School of Electronic Engineering
Dublin City University
Glasnevin, Dublin 9
Email: barryl@eeng.dcu.ie

***Abstract*—The realisation of cost-efficient Optical Burst Switching (OBS) networks can be greatly facilitated from minimising the number of contention resolution resources required at congested network nodes. In this paper we present a Fibre Delay Line (FDL) optimal allocation scheme where the total cost associated to the employment of FDLs is minimised subject to performance requirements defined in terms of maximum tolerable end-to-end blocking probability. The optimal buffer configuration is achieved by means of a constraint-handling genetic algorithm. We additionally increase the accuracy of our analysis by considering the non-Poissonian traffic characteristics of the OBS network under study. Results show that our method permits to identify an optimal FDL configuration that minimises the total buffer installation cost and simultaneously satisfies the network blocking probability requirements.**

***Keywords*-Optical Burst Switching; Fibre Delay Lines; Genetic Algorithms; Optimisation;**

I. Introduction

In recent years substantial research effort has been devoted to the performance evaluation of network architectures employing Optical Burst Switching [1], one of the most promising switching strategies for the deployment of next generation optical networks. A major drawback in OBS is due to burst loss which occurs when two packets (*bursts*) are contending for the same wavelength channel on a common output fibre link. This issue can be addressed with the employment of Fibre Delay Lines (FDLs) [1], [2]. An FDL can be considered as a buffer in the time domain and is capable of preventing burst loss by delaying the transmission of one of the contending bursts. It has been demonstrated that FDLs can be very effective in reducing burst loss of several orders of magnitude as shown in works such as [2], [3] and [4], where performance evaluations of buffered OBS architectures have been conducted. The overall performance of an FDL-buffered OBS network might vary considerably depending on how many FDLs are employed and their allocation in the network. In non-uniform network topologies some links of the network may be congested more than others even under uniform end-to-end traffic demands. This means that some links may require more buffering resources than others resulting in a non-uniform allocation of FDLs in the network, however, the problem can not be solved by simply adding FDL buffers to bottleneck links. In fact, the employment of an FDL might shift the traffic load from a congested link to the next link over the same path, thus potentially shifting the "congestion problem".

In this paper we address this issue by proposing a method to find an optimal FDL allocation that minimises the cost associated to the buffers employment and, at the same time, satisfies a maximum tolerable end-to-end blocking probability. We solve this problem by means of genetic algorithms [5], a branch of evolutionary algorithms that have been already successfully used to solve different optimisation problems for photonic switched networks. For example, in [6] the authors develop a genetic algorithm to jointly solve a Routing and Wavelength Assignment (RWA) problem for optical networks. A similar method has been derived in [7], where the authors solve an RWA problem for Optical Packet Switching (OPS) networks with load balancing. Yang et al. propose in [8] a multi-objective genetic algorithm to simultaneously minimise the delay while maximising the throughput for metro optical networks. Differently from these works, our main contribution is in applying a genetic algorithm to solve a cost minimisation problem where the decision variables define the *allocation of the FDL wavelength channels*. Castro et al. focus on a similar problem in [9] where they derive a method to find an optimal placement of the FDLs in OBS networks with Tabu Search, however, differently from [9], we decide to use a different OBS node buffered architecture [3] and a more realistic and accurate network model as proposed in [10]. The rest of the paper is organised as follows: in Section II, we briefly describe the architecture and the analytic model of the OBS network under study. In Section III, we define a cost minimisation problem for the OBS network in question and in Section IV we describe the genetic algorithm used to solve it. Results and conclusions are respectively given in Section V and VI.

II. The OBS Network Under Study

We consider the Tune And Select with Shared feedback FDL (TAS-shFDL) OBS node architecture analysed in [3] and illustrated in Figure 1(a). The switch is equipped with

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

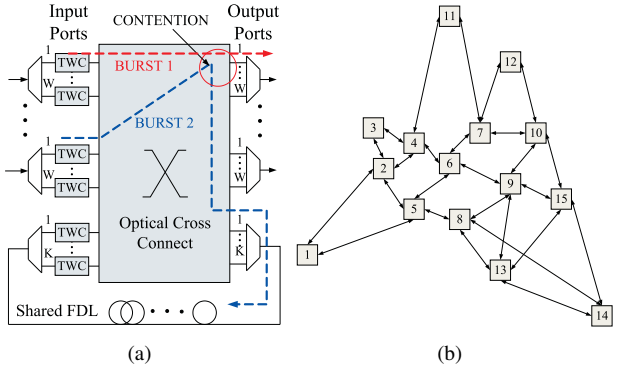


Figure 1. The architecture of the buffered OBS node (a) and the OBS European Optical Network (EON) topology under study (b).

P input/output ports, each one connected to an optical fibre link comprising W wavelength channels. We assume full wavelength conversion, that is each channel is supported by a Tunable Wavelength Converter (TWC) for burst contention resolution. Additionally, an extra input/output port is dedicated to an FDL comprising K wavelength channels. We refer to these channels as *virtual buffers* as described in [2]. The FDL is shared between the output links connected to the node in a feedback configuration [3]. This means that a contention between two bursts will be resolved by directing one of the bursts to a free virtual buffer of the FDL and then re-offering it to a free wavelength channel of the output port. If this is not possible, the burst will be dropped and consequently lost from the system.

We consider an OBS network of such switches described by a graph $\mathcal{G}(N, L, R)$, where N is the number of nodes, L is the number of links and R is the number of paths of the network. All links comprise the same number of wavelength channels W. Each path r is offered with burst traffic of load ρ_r (in Erlang). We further define $\boldsymbol{\rho}= [\rho_1, \rho_2, \ldots, \rho_R]$ as the vector comprising the burst traffic loads offered to each path. We characterise the burst traffic as a non-Poisson process by assuming generally distributed burst interarrival times and exponentially distributed burst lengths. We attempt to model the traffic characteristics with the BPP two-moment matching technique [11] by considering the additional contribution of the traffic *peakedness* Z. The peakedness quantifies the deviation of the burst traffic from being Poisson and is defined as the ratio between the variance and the load of the burst traffic. The traffic is said to be *peaked* or *smooth* whether Z is greater or less than one. If $Z = 1$ the traffic is Poisson. This analysis allows us to approximately match the expected OBS traffic characteristics, which are largely determined by the burst aggregation process [12]. Further results on the impact of traffic burstiness in optical packet switching networks can be found in [13]. Under these premises, we model the OBS network with the approximate method proposed by the present authors in [10]. The model is used to evaluate end-to-end burst blocking probabilities and can generally be summarised as a non linear function whose output is the vector $\boldsymbol{\mathcal{P}}= [\mathcal{P}_1, \ldots \mathcal{P}_R]$, where $\mathcal{P}_r$ is the end-to-end blocking probability of path r. Namely,

$$\boldsymbol{\mathcal{P}} = \boldsymbol{\mathcal{P}}(N, L, R, W, \mathbf{K}, \boldsymbol{\rho}, \mathbf{Z}), \quad (1)$$

where we have indicated with $\mathbf{Z} = [Z_1, \ldots Z_R]$ the vector of the burst traffic peakednesses offered to each path and with $\mathbf{K} = [K_1, \ldots, K_N]$ the vector comprising the number of virtual buffers allocated to each node in the network. For space constraints we can not provide a description of the method in this paper. The reader will find a detailed mathematical analysis and the validation of our method in [10], however we show here some additional new results in Figures 2 and 3. Particularly, the average end-to-end burst blocking probability obtained with our analytic model is compared with the one obtained from a discrete-event simulation of the OBS European Optical Network (EON) topology depicted in Figure 1(b). The network comprises $N = 15$ nodes, $L = 25$ bidirectional links and $R = 18$ source-destination pairs whose shortest paths are indicated in Table I. The FDL allocation is uniform, thus all nodes are equipped with the same number of virtual buffers. The traffic demands are uniform as well, that is each path is offered with the same traffic load and peakedness. As we can see from the graphs the accuracy of the analytic model compares favourably with the simulation data for a broad range of end-to-end blocking probability, a feature that convinces us to adopt our model for the definition and the solution of the cost minimisation problem.

III. Definition of the Optimisation Problem

We first start by introducing the cost function that will be used to define the objective of the optimisation problem. Our goal is to determine an estimate of the cost introduced by the employment of a shared FDL in a node of the network. Following the analysis presented in [3] on the

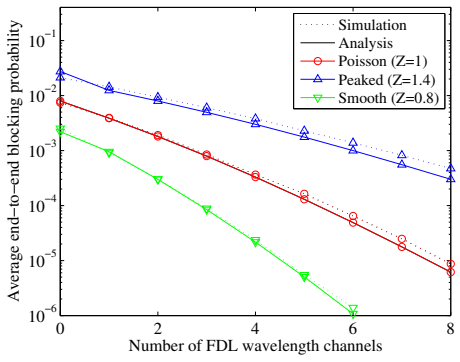


Figure 2. Average end-to-end blocking probability vs. number of FDL channels for the EON topology. The number of wavelength channels per link is $W = 16$ and the normalised load per path is $\rho_r = 0.25$ Erlang.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

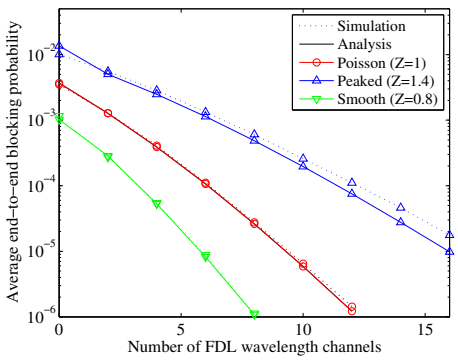


Figure 3. Average end-to-end blocking probability vs. number of FDL channels for the EON topology. The number of wavelength channels per link is $W = 32$ and the normalised load per path is $\rho_r = 0.3$ Erlang.

TAS-shFDL architecture, we note that the installation of an extra input/output port dedicated to the FDL requires one additional Erbium Doped Fibre Amplifier (EDFA). Furthermore, since we are assuming full wavelength conversion, each wavelength channel of the FDL must employ a TWC, for a total of K_n TWCs. Finally, in order to allow the transmission of burst packets to the FDL, each wavelength channel on each output port must be equipped with an additional Semiconductor Optical Amplifier (SOA), for a total of $P_n \cdot W$ SOAs, where we have indicated with P_n the number of output ports of node n. Similarly, in order to send packets to the output ports, each wavelength channel of the FDL requires P_n SOAs for a total of $P_n \cdot K_n$ SOAs. Under these premises, we define the total cost associated with an FDL to node n as follows,

$$C_n = h_E + h_T K_n + h_S P_n (W + K_n), \tag{2}$$

where we have denoted with h_E, h_T and h_S respectively the unit cost of an EDFA, of a TWC and of a SOA. Finally, the total cost arising from the employment of FDLs in the network can be expressed as

$$C(\mathbf{K}) = \sum_{n=1}^{N} C_n = \sum_{n=1}^{N} \left[h_E + h_T K_n + h_S P_n (W + K_n) \right], \tag{3}$$

Table I
PATHS OF THE EUROPEAN OPTICAL NETWORK TOPOLOGY.

Path	Path hops	Path	Path Hops
1	$1 \to 2 \to 4 \to 6 \to 7 \to 10$	10	$11 \to 7 \to 12$
2	$3 \to 4 \to 6$	11	$12 \to 10 \to 15 \to 14$
3	$13 \to 15 \to 10 \to 12$	12	$10 \to 7 \to 11$
4	$12 \to 7 \to 6 \to 4 \to 2$	13	$13 \to 9 \to 6 \to 4 \to 11$
5	$2 \to 4 \to 11$	14	$8 \to 5 \to 2 \to 3$
6	$11 \to 7 \to 6 \to 5 \to 8$	15	$4 \to 2 \to 1$
7	$12 \to 10 \to 9 \to 13$	16	$7 \to 10 \to 15$
8	$5 \to 8 \to 13 \to 14$	17	$13 \to 8 \to 5 \to 1$
9	$1 \to 5 \to 6 \to 7$	18	$14 \to 15 \to 10 \to 7 \to 11$

where $\mathbf{K} = [K_1, \dots K_N]$ is a vector representing the FDLs allocation in the network. We are now ready to define the following problem:

Given an OBS network defined by graph $\mathcal{G}(N, L, R)$ where each link comprises the same number of channels W and where the traffic demands for each path are quantified by vectors $\boldsymbol{\rho}$ and $\mathbf{Z}$, we want to minimise the cost function C as follows,

$$\begin{aligned} \underset{\mathbf{K}}{\text{minimise}} \quad & C(\mathbf{K}) \\ \text{subject to} \quad & \mathcal{P}_r(\mathbf{K}) \leq \mathcal{P}_{max}, && r = 1, \dots, R, \\ & K_n \leq K_{max}, \quad K_n \in \mathbb{N}, && n = 1, \dots, N, \end{aligned} \tag{4}$$

where we have indicated with $\mathcal{P}_{max}$ the maximum tolerable end-to-end blocking probability and with K_{max} the maximum number of virtual buffers that can be allocated in a node of the network. Furthermore, we force the number of virtual buffers to be positive integers. We solve the above defined problem with the use of a genetic algorithm as described in the next section.

IV. GENETIC ALGORITHM

Genetic algorithms (GAs) [5] are a branch of *evolutionary* algorithms, a family of search heuristics that mimics the process of evolution to find near-optimal solutions for optimisation problems. In a GA, each potential solution corresponds to a string of decision variables called an *individual* (or *chromosome*) where each decision variable represents a *gene*. The algorithm starts by generating an initial random population of individuals. A set of individuals is selected from the population to form a new generation on the basis on "how suitable" they are as solutions of the optimisation problem. The "goodness" of the selected individuals is evaluated by a specific *fitness* function which is typically defined as a combination of the objective functions of the optimisation problem in question. In this way, the better individuals have more chances to "reproduce" and transfer their "good" genes to their children (*offspring*) that will form a better new generation, mimicking the evolution process. The algorithm normally ends when a user-defined maximum number of generations is reached or when some conditions on the improvement achieved by the best individuals are met.

A. Initial Population and Encoding

In our problem, each individual corresponds to a specific allocation of FDLs $\mathbf{K}$. Each element of $\mathbf{K}$ is the number of wavelength channels of an FDL at a given node and represents a gene of the individual. All the individuals are encoded directly into strings of integer numbers with values in the range $[0, K_{max}]$. Note that the encoding process forces the potential solutions to be integrals and within the interval $[0, K_{max}]$. Therefore, the constraints $K_n \leq K_{max}$ and $K_n \in \mathbb{N}$ for $n = 1, \dots, N$ are already satisfied by the process of encoding of the individuals.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

B. Fitness function and Selection

The fitness function evaluates the "goodness" of an individual. The greater is the fitness value of an individual, the higher is the probability that the individual will be selected for "reproduction". Generally, in constrained optimisation problems, the fitness function of each individual is modified by introducing a non-zero *penalty function* for the solutions that are *unfeasible*, that is the solutions that do not satisfy the constraints of the optimisation problem. We adopt a simple yet very efficient method inspired by the work of Deb in [14]. Particularly, the fitness function f of an individual $\mathbf{K}$ can be written as

$$f(\mathbf{K}) = \begin{cases} -C(\mathbf{K}) & \text{if } \mathbf{K} \text{ is feasible,} \\ -C(\mathbf{K}^-) - |\mathcal{P}_r(\mathbf{K}) - \mathcal{P}_{max}| & \text{if } \mathbf{K} \text{ is unfeasible,} \end{cases} \quad (5)$$

where we have indicated with $\mathbf{K}^-$ the feasible FDL allocation with the lowest fitness in the population and with $|\mathcal{P}_r(\mathbf{K}) - \mathcal{P}_{max}|$ the *constraint violation* of individual $\mathbf{K}$ representing the penalty function for $r = 1, \ldots, R$. At each generation, the fitness of all individuals is evaluated and a set of "good" candidate solutions are selected to "reproduce". The *selection* process is a key operation in genetic algorithms and there are several mechanisms to perform it. We decide to select individuals with the *roulette wheel technique* [5] where the fittest individuals have more chances to be chosen for reproduction. Particularly we first normalise the fitness value of all the individuals of the population as

$$f_i^* = f_i / \sum_{i=1}^{I} f_i \qquad i = 1, \ldots, I, \quad (6)$$

where I is the number of individuals in the population and f_i is the fitness of individual i. Then, we sort the fitness values in ascending order (denoting them with t_i^*) and we generate a random number ϵ uniformly distributed within the interval [0,1]. If $\epsilon < t_1^*$, we select individual 1 as a parent for reproduction. If $\epsilon > t_1^*$, we calculate the cumulative sum $s_1 = t_1^* + t_2^*$ and we compare again ϵ with s_1. At this point, if $\epsilon < s_1$, we select individual 2 otherwise we recursively re-calculate the cumulative sum $s_2 = s_1 + t_3^*$ and proceed with the next comparison in a similar manner until two individuals will be selected as parents.

C. Crossover, Mutation and Elitism

Once the individuals have been selected, they reproduce to generate a new offspring. This step of the algorithm is called *crossover* and is performed with a user-defined probability $Prob_c$. We choose to perform a *two-point crossover* where the new offspring inherits genes from the parents on the basis of two random crossover points as illustrated in Figure 4.

Once the new children are generated, we *mutate* them by randomly changing one of their genes with a predefined mutation probability $Prob_m$. The mutation is an essential step in GAs that helps preserving the diversity in the population and prevents the GA to get stuck in a local minimum.

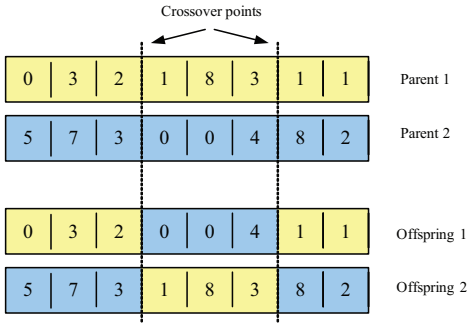


Figure 4. Example of crossover. In this case each individual is encoded as a string of 8 integers.

In the final step of the algorithm, once the new generation is obtained, we select a specific number of individuals E with the highest values of fitness and we include them in the new generation. This final step is known as *elitism* and the set of chosen individuals is called the *elite*. This procedure permits us to keep the best E individuals in the population as the algorithm continues its search for fitter solutions.

V. Results

We test our method on the same network topology of Figure 1(b). The configuration settings for the genetic algorithm are shown in Table II. The values of $Prob_c$ and $Prob_m$ are proved to generally work well for different optimisation problems. The estimation of the hardware unit costs h_S, h_E and h_T is quite difficult as real costs for these devices vary considerably on the basis of their manufacturer and their specifications. Based on the study proposed in papers such as [15] and [16] we decided that it may be reasonable to relate all unit costs to the one of a SOA, being the SOA a device less expensive than an EDFA and a TWC. Thus, we set the unit cost of a SOA as h_S = 1 and we decide to fix the unit cost of an EDFA at $3h_S$ and the unit cost of a TWC at $15h_S$. We stop the genetic algorithm after 300 generations. Table III and Table IV illustrate the benefits introduced by the optimisation in terms of cost savings subject to different values of $\mathcal{P}_{max}$ for different values of traffic load and peakedness. We first note how

Table II
Genetic Algorithm parameters configuration

Population Size	80
Elite Size (E)	16
Selection	Roulette Wheel
Crossover	Two-point
$Prob_c$	0.9
$Prob_m$	0.05

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Table III
COST COMPARISON BETWEEN OPTIMAL (OPT) AND UNIFORM (UNI) VIRTUAL BUFFER ALLOCATION FOR $K_{max} = 8$ BUFFERS, $W = 16$, $\rho = 0.3$ ERLANG FOR EACH PATH. 'NF' STANDS FOR 'NOT FEASIBLE'.

	$\mathcal{P}_{max}$	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Z = 0.8	C_{OPT}	0	601	1209	1620	NF
	C_{UNI}	0	1179	1968	2757	NF
Z = 1	C_{OPT}	0	1020	1620	NF	NF
	C_{UNI}	0	1705	2757	NF	NF
Z = 1.4	C_{OPT}	87	1540	NF	NF	NF
	C_{UNI}	1179	2757	NF	NF	NF

Table IV
COST COMPARISON BETWEEN OPTIMAL (OPT) AND UNIFORM (UNI) VIRTUAL BUFFER ALLOCATION FOR $K_{max} = 16$ BUFFERS, $W = 32$, $\rho = 0.35$ ERLANG FOR EACH PATH. 'NF' STANDS FOR 'NOT FEASIBLE'.

	$\mathcal{P}_{max}$	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Z = 0.8	C_{OPT}	0	731	1937	2617	3065
	C_{UNI}	0	2313	3365	4154	4680
Z = 1	C_{OPT}	0	1536	2477	3210	NF
	C_{UNI}	0	3102	3891	4943	NF
Z = 1.4	C_{OPT}	0	2178	3479	NF	NF
	C_{UNI}	0	3891	5469	NF	NF

the cost varies considerably with Z, an occurence that, we believe, justifies the choice of modelling the OBS network with the analytic method proposed in [10]. We compare the total FDL cost resulting from our optimisation method (C_{OPT}) with the total cost resulting from the minimum uniform allocation of the virtual buffers that satisfies the requirements in terms of $\mathcal{P}_{max}$ (C_{UNI}). For example, in Table IV, to reach a maximum target blocking probability $\mathcal{P}_{max}$ of 10^{-2} on all paths for $Z = 1.4$, the optimal numbers of FDLs are found to be $\mathbf{K}_{opt}$=[0 8 0 10 4 6 10 4 0 10 10 4 0 0 4], resulting in a total cost of $C_{OPT} = 2178$. The same performance requirements can be satisfied with a uniform allocation of no less than 10 buffers in each node, for a total cost of $C_{UNI} = 3891$. Thus, for this particular scenario, the optimisation process yields a 44% reduction in cost of the extra hardware added by the employment of FDLs compared to a uniform FDL allocation. Furthermore, following [3], we can also estimate the achieved reduction in the total hardware cost of the network with the same optimal buffer allocation. In fact, in the bufferless TAS OBS node architecture (that is, without considering the extra hardware added by the FDL), a node n is equipped with $2P_n$ EDFAs, $W \cdot P_n$ TWCs and $W \cdot P_n^2$ SOAs. If we consider the cost of this additional hardware in the OBS network under study for all nodes of the same scenario above mentioned, we obtain a total hardware network cost of $C_{OPT} = 24294$ for the optimal allocation $\mathbf{K}_{opt}$ and $C_{UNI} = 26007$ for the uniform allocation of 10 buffers per node, resulting in an approximate total hardware cost saving percentage of 6.6%. We also note that for some scenarios it is not possible to find an optimal (and uniform) allocation of the FDLs (e.g., Table III for $Z = 1.4$ and $\mathcal{P}_{max} = 10^{-3}$). This is because all the solutions found are unfeasible, that is there is no FDL allocation that can satisfy the performance requirements given by $\mathcal{P}_r(\mathbf{K}) \leq \mathcal{P}_{max}$ for $r = 1, \ldots, R$ with $K_n \leq K_{max}$ for all nodes of the network.

Figure 5 illustrates an example of the distribution of the FDL virtual buffers in the OBS network. We observe that the FDL distribution changes considerably with Z, since congestion at nodes increases when traffic becomes peaked. Note that some nodes are not assigned with FDLs, regardless of the peakedness of their offered traffic demands. Thus, the genetic algorithm is able to identify the nodes of the network for which adding an FDL does not add any contribution in lowering the end-to-end blocking probability value. In this regard we want to remark that, although the offered load may be generally considered low in all the cases of study (0.3-0.35 Erlang), this is not the case for congested links in the core network, where it can reach normalised values of 0.6 Erlang. The algorithm allows to determine the optimal number of FDL buffers required for nodes with such congested links, a number that is higher than the one determined for the less congested links at the edges of the network. This feature may consent to considerably decrease the FDL cost compared to an uniform allocation as shown in the example of Figure 6. In this particular case, to satisfy the performance requirements, at least 4 FDL buffers must be employed to node 7. This means that, in an uniform allocation, we must employ at least 4 FDL buffers for each node of the network, resulting in an increased FDL cost per network node compared to the optimal scenario.

Finally, end-to-end blocking probabilities for each path are shown in Figure 7. We observe that the analytic method provides a quite accurate estimate of the blocking probability at the optimal point compared to simulation data. The graph

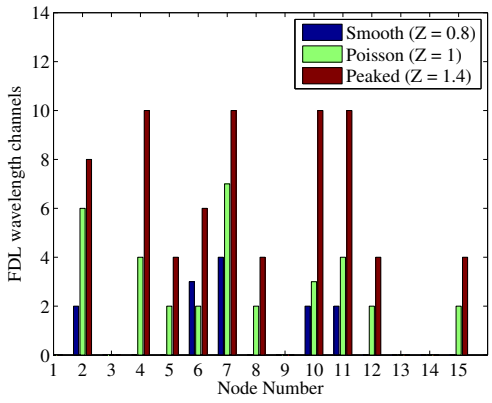


Figure 5. Allocation of the FDLs in the network for $W = 32$,$\rho = 0.35$ Erlang and $\mathcal{P}_{max} = 10^{-2}$. Note that nodes 1,3,9,13 and 14 do not contribute in lowering the blocking if equipped with FDLs, thus they are not assigned with FDLs.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

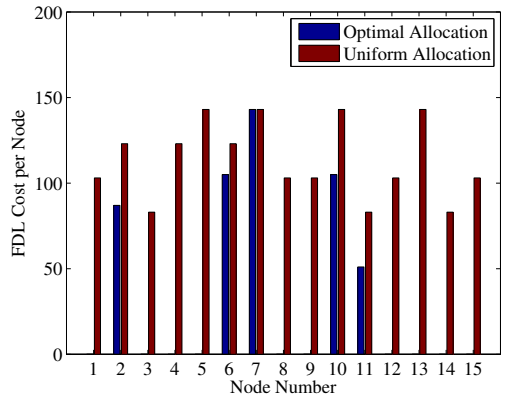


Figure 6. Comparison of the FDL cost per node between optimal and uniform FDL allocation for $W = 32$,$\rho = 0.35$ Erlang and $Z = 0.8$ and $\mathcal{P}_{max} = 10^{-2}$. The optimal FDL allocation is found to be **K**=[0 2 0 0 0 3 4 0 0 2 2 0 0 0 0]. The uniform allocation forces each node to employ at least 4 FDL buffers, that is the required minimum number of FDL buffers to satisfy the performance constraints at node 7.

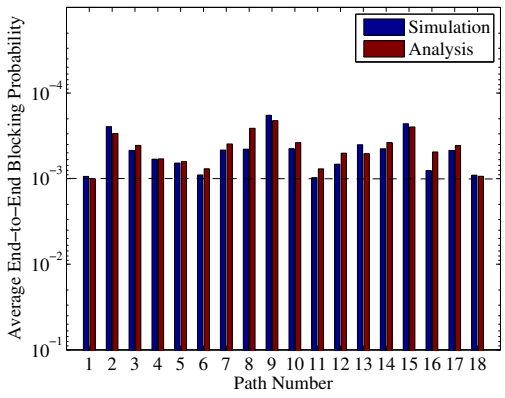


Figure 7. Path blocking probabilities for $W = 16$,$\rho = 0.3$ Erlang and $Z = 1$. Note that all blocking values compare favourably with simulation results and they are all below the required performance level given by $\mathcal{P}_{max} = 10^{-3}$. The optimal FDL allocation for this particular scenario is found to be **K**=[0 6 0 7 6 7 8 6 0 6 6 6 0 0 6].

additionally shows that each path blocking is below the maximum tolerable value given by $\mathcal{P}_{max}$, thus satisfying the performance constraint of our optimisation problem.

VI. Conclusions

We have proposed a method to find an optimal allocation of FDLs in an OBS network that minimises the cost associated with the employment of FDL-buffers and satisfies performance requirements in terms of maximum tolerable end-to-end blocking probability. Our results illustrate the potential equipment cost savings achieved when FDL allocation is optimised as opposed to uniformly distributing the number of buffers in the network. Future works will deal with the definition of multi-objective optimisation problems for OBS networks where conflicting objectives such as throughput maximisation and cost minimisation will be taken into consideration.

Acknowledgment

This material is based on research supported by Science Foundation Ireland (SFI) under the Research Frontiers Programme Grant No. [08/RFP/CMS1402].

References

[1] M. Yong, C. Qiao, and S. Dixit, "QoS Performance of Optical Burst Switching in IP-over-WDM Networks," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 10, pp. 2062-2071, October 2000.

[2] X. Lu and B. L. Mark, "Performance Modeling of Optical-Burst Switching with Fiber Delay Lines," *IEEE Trans. Commun.*, vol. 52, no. 12, pp. 2175-2183, December 2004.

[3] C. M. Gauger, H. Buchta, and E. Patzak, "Integrated Evaluation of Performance and Technology - Throughput of Optical Burst Switching Nodes Under Dynamic Traffic," *IEEE J. Lightw. Technol.*, vol. 26, no. 13, pp. 1969-1979, July 2008.

[4] W. Rogiest, D. Fiems, K. Laevens, and H. Bruneel, "Modeling the performance of FDL buffers with wavelength conversion," *IEEE Trans. Commun.*, vol. 57, pp. 3703-3711, December 2009.

[5] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Publishing Company, 1989.

[6] I. de Miguel, R. Vallejos, A. Beghelli, and R. Duran, "Genetic Algorithm for Joint Routing and Dimensioning of Dynamic WDM Networks ," *IEEE J. Opt. Commun. Netw.*, vol. 1, no. 7, pp. 608-621, December 2009.

[7] H. Tode, K. Hamada, and K. Murakami, "ORGAN: Online route and Wavelength design based on Genetic Algorithm for OPS networks," *Proc. of ONDM 2010*, pp. 1-6, February 2010.

[8] H. Yang, M. Maier, M. Reisslein, and W. M. Carlyle, "A Genetic Algorithm-Based Methodology for Optimizing Multiservice Convergence in a Metro WDM Network," *IEEE J. Lightw. Technol.*, vol. 21, no. 5, pp. 1114-1133, July 2003.

[9] J. C. S. Castro, J. M. F. Pedro, and P. P. Monteiro, "Routing and Buffer Placement Optimization in Optical Burst Switched Networks," *Proc. of IEEE CLCN 2008*, pp. 353-360, 14-17 Oct. 2008.

[10] D. Tafani, C. McArdle, and L. P. Barry, "Analytical Model of Optical Burst Switched Networks with Share-per-Node Buffers," *Proc. of ISCC 2011*, pp. 512-518, Corfu, 28 June - 01 July 2011.

[11] L.E.N. Delbrouck, "The Uses of Kosten's System in the Provisioning of Alternate Trunk Groups Carrying Heterogeneous Traffic," *IEEE Trans. Commun.*, vol. COM-31, no. 2, pp. 741-749, 1983.

[12] A. Rostami and A. Wolisz, "Modeling and Synthesis of Traffic in Optical Burst-Switched Networks," *IEEE J. Lightw. Technol.*, vol. 25, no. 10, pp. 2942-2952, October 2007.

[13] H. Øverby and N. Stol, "Effects of Bursty Traffic in Service Differentiated Optical Packet Switched Networks," *Optics Express*, vol. 12, no. 3, pp. 410-415, 2004.

[14] K. Deb, "An Efficient Constraint Handling Method for Genetic Algorithms," *Computer Methods in Applied Mechanics and Engineering*, vol. 186, no. 2-4, pp. 311-338, June 2000.

[15] V. Eramo, M. Listanti, and A. Germoni , "Cost Evaluation of Optical Packet Switches Equipped With Limited-Range and Full-Range Converters for Contention Resolution," *IEEE J. Lightw. Technol.*, vol. 26, no. 4, pp. 390-407, 2008.

[16] C. Raffaelli, and M. Savi, "Cost Comparison of All-Optical Packet Switches with Shared Wavelength Converters," *Proceedings of ICTON 2007*, vol. 3, pp. 209-212, July 2007.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Outage Capacity of Mobile Wireless Optical Link in Indoor Environment

Nicolas Barbot, Seyed Sina Torkestani, Stephanie Sahuguede, Anne Julien-Vergonjanne, Jean-Pierre Cances
XLIM DPT-C2S2 UMR CNRS 6172/ ENSIL, 16 rue d'Atlantis, 87068 LIMOGES, FRANCE
{nicolas.barbot, seyed.torkestani, s_sahuguede, anne, cances}@ensil.unilim.fr

***Abstract*—In this paper, we investigate the potentiality of a wireless optical communication system in an indoor environment for both line of sight (LOS) and diffuse configurations by taking into account transmitter mobility. Statistical model of the mobile optical channel is defined for each configuration. Our contribution is to determine the outage probability and then the outage capacity considering an On-Off Keying modulation and different data rates. From the results one can obtain the outage probability value which maximizes the outage capacity for the considered indoor mobility scenario. We finally show the LOS performance gain provided by Forward Error Correction (FEC), considering Low-Density Parity-Check (LDPC) codes of different lengths.**

***Index Terms*—Indoor Wireless Optical Communications, LOS Links, Diffuse Links, Outage probability, Outage Capacity**

I. Introduction

Nowadays, wireless optical communications are popular technologies which offer many advantages such as low complexity implementation and high secured transmissions. Infrared (IR) transmissions constitute an interesting alternative solution to radio-frequency (RF) one for many indoor and home applications [1]. Actually, IR systems intrinsically offer several benefits over RF systems due to the absence of multipath fading and robustness to electromagnetic interferences (EMI). Moreover, optical systems permit having an unregulated and quasi-unlimited bandwidth [1,2]. However, IR systems suffer from a high path loss, a reduced coverage area and lower link budget compared to RF systems.

Two kinds of IR systems are generally considered: LOS (Line Of Sight) links and diffuse links [2]. The LOS propagation is the most commonly used scheme. It permits guaranteeing high Signal to Noise Ratio (SNR). For IR short range indoor communications, the main drawback is the severe impact on the path loss due to misalignments between transmitter and receiver [1]. On the other side, in diffuse configuration, the light is emitted toward a reflective surface and the detector collects the reflected power from this surface. Diffuse configuration permits obtaining higher coverage area than in the LOS one but this is done at the cost of a reduction of the optical received power (lower SNR).

However, for both systems, mobility of transmitter or receiver significantly decreases the performances due to the variations of the distance between transmitter and receiver. In this case, outage probability evaluation permits quantifying attainable data rates and quality of service for the optical communicating system. Besides, the mobile system potentiality can be well described by analyzing the channel capacity.

The paper is organized as follows: after presenting the state of the art in Section II, the optical transmission system is described in Section III. We then evaluate the outage probability in Section IV in LOS and diffuse configurations using a statistical approach. In Section V, we estimate the outage capacity for both configurations considering the non-stationary channel. In order to illustrate the performance of a FEC, Section VI deals with different LDPC codes in a LOS configuration.

II. State of the art

Capacity of optical channel has been already studied in the case of free space optical transmissions over atmospheric channel subject to scintillation in [5] and [6], respectively, with and without channel side information. Capacities of other outdoor optical channel have been explored, in [7], authors analyse the effect of pointing errors and in [8] and [9], authors determine the capacity considering multiple receivers. However, to the best of our knowledge, capacity of indoor mobile optical channel has not been yet investigated.

Our contribution is to determine the performance of LOS and diffuse configurations considering a statistical model of the mobile indoor optical channel to evaluate the outage probability. From the outage probability analysis, we evaluate the capacity of this channel for both LOS and diffuse configurations. Besides, we illustrate the gain provided by Forward Error Correction (FEC) on the LOS optical transmission by considering different LDPC codes.

III. System Description

We consider an indoor environment and a communication link between a mobile transmitter in the environment and a base station placed on the ceiling.

The transmitter is at (x_1, y_1, z_1) in a room supposed to be free of any obstacles and represented by a box of dimensions (3m,4m,2.5m). The receiver is supposed to be placed on the middle of the ceiling at ($x_2 = 1.5$m, $y_2 = 2$m, $z_2 = 2.5$m) and is pointed toward the floor to achieve minimum path losses (see Fig. 1).

Data are sent by using an IR communication system based on Intensity Modulation and Direct Detection (IM/DD). The transmitted signal is thus an optical power which is always

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

positive and the channel can be modeled by a linear system [1]. The received signal $Y(t)$ can be written as:

$$Y(t) = RX(t) \otimes h(t) + N(t) \quad (1)$$

where $X(t)$ is the instantaneous optical power, R is the photo-diode responsivity, and $h(t)$ represents the impulse response of the optical channel. $N(t)$ represents the Additive White Gaussian Noise (AWGN) [3].

In the following, we study two IR propagation types: LOS and diffuse ones. For LOS case, the directivity of both optical emitter and receiver does not allow multipath propagation. For diffuse case, delay spread D is typically equal to 10 ns [1] and is supposed to be negligible compared to low rate transmission ($D << 1/R_b$). For higher rates, intersymbol interference can be compensated by an equalization module. Thus the impulse response is only characterized by its static gain H such as: $h(t) = H\delta(t)$. On Off Keying (OOK) modulation is used to transmit symbols over the AWGN channel. At the reception, the electrical SNR is proportional to the square of the received optical power due to photo-diode detection [1]:

$$SNR = \frac{2R^2P_t^2H^2}{N_0R_b} \quad (2)$$

where P_t is the average transmitted power, N_0, the noise power-spectral density and R_b the transmission data rate.

In this study, we have chosen $R = 0.55$ A/W. N_0 is determined considering that shot noise is the dominant noise source [2]: $N_0 = 2I_bq$ with mean current $I_b = 200$ μA and $q = 1.6 \times 10^{-19}$ C thus $N_0 = 6.4 \times 10^{-23}$ W/Hz.

A difference between LOS and diffuse configurations appears in H expression. In a LOS configuration, the static gain H directly depends on the distance d between the transmitter and the receiver and can be evaluated by [1]:

$$H = \frac{A}{\pi d^2} \quad (3)$$

where A is the photo-detector physical surface. Note that this corresponds to the case where LOS transmitter is perfectly aligned with the receiver.

For the diffuse configuration, the channel gain is obtained using ceiling bounce model [4]. The received power is computed by summing all the contributions of tiny elements of the reflective surface (the floor). The static gain can be thus expressed by:

$$H = \frac{\rho A z_1^2 z_2^2}{\pi^2} \times \iint_{plan} \frac{\mathrm{d}x\mathrm{d}y}{(z_1^2+(x-x_1)^2+(y-y_1)^2)^2(z_2^2+(x-x_2)^2+(y-y_2)^2)^2} \quad (4)$$

where ρ is the floor reflectivity.

For both LOS and diffuse configurations, we consider a typical physical area $A = 1$ cm^2 and $\rho = 0.8$. In order to respect eye safety regulations, P_t have been set to 20 mW for the LOS configuration and to 300 mW for the diffuse one which are the typical allowed transmitted power [10]. The diffuse Field Of View (FOV) of the receiver is set to 70°.

In order to represent the transmitter mobility, as a first approach, we model its location within the room by Gaussian distributions in x axis from 0 to 3m, in y axis from 0 to 4m and in z axis from 0 to 1.5m with respectively $\mathcal{N}(1.5, 0.25)$ and $\mathcal{N}(2, 0.36)$ and $\mathcal{N}(1.2, 0.09)$ distributions. The means of the position distributions along x and y axis are chosen equal to the middle of the room, which means that the transmitter presence is more probable in this area (for example for a transmitter placed on a person who moves inside the room). The mean of the distribution in z is chosen equal to 1.2 m and corresponds for example to a transmitter placed at the belt level of a person. The variances of the distributions were defined so that to include 98% of distribution data inside the room.

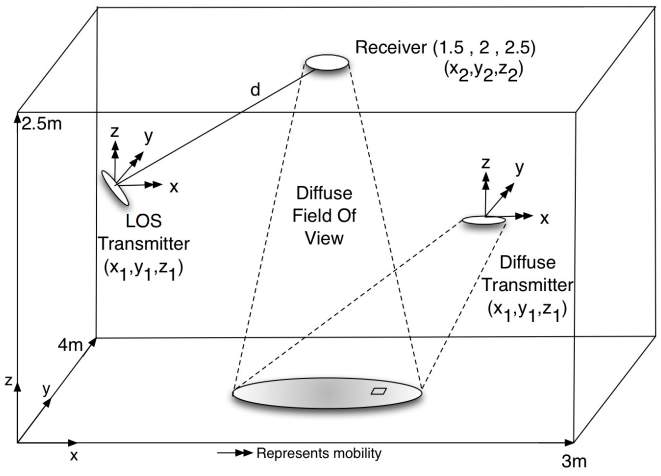


Fig. 1. Room Configuration

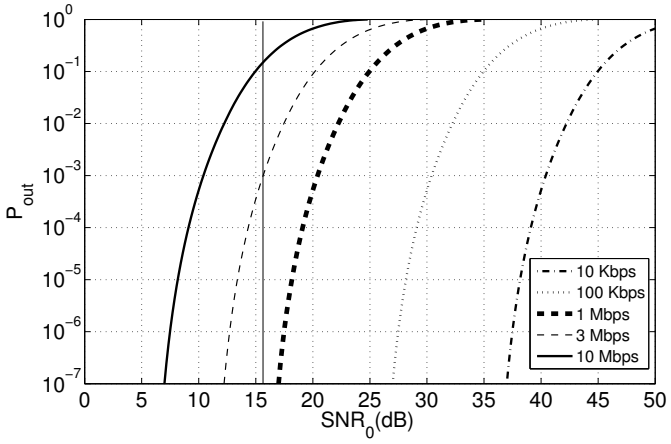


Fig. 2. Outage Probability versus SNR in LOS configuration for OOK modulation

We can note that by considering mobility, H varies due to the distance variations between the transmitter and the receiver. This is analyzed in the following Sections.

IV. Outage Probability

In the context studied, H variations are slow in the bit time even for the lowest considered data rate. Optical channel can

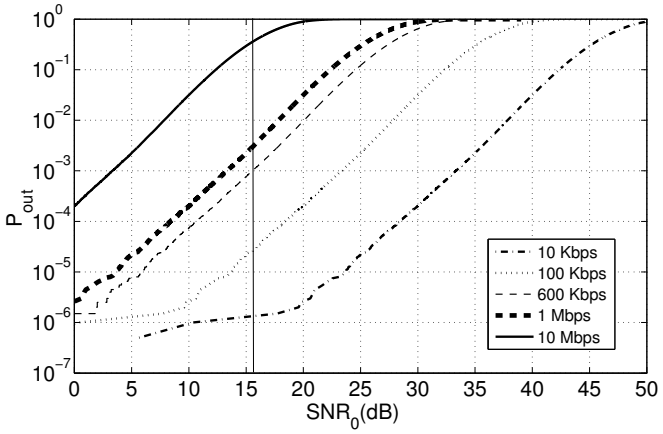


Fig. 3. Outage Probability versus SNR in diffuse configuration for OOK modulation

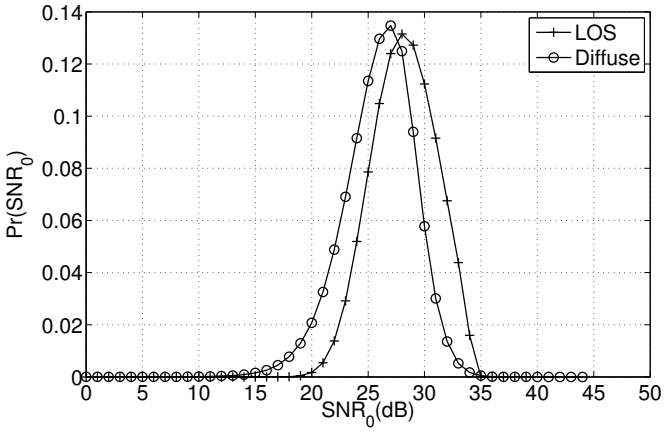


Fig. 4. Distribution of SNR for $R_b = 1$Mbps

be thus considered as a slow fading channel [1]. Consequently, average BER does not represent a good metric to describe transmission performance. Instead, the outage probability is used to estimate the performance. The outage probability is defined as the probability that the capacity C of the channel does not support the rate R_0 of the transmission [11]. Since the capacity is a monotonic function of SNR, the outage probability can be expressed as a function of a SNR threshold (SNR_0) and corresponds to the probability that the SNR value at a given time drops below SNR_0:

$$P_{out} = \Pr[C(SNR) < R_0] = \Pr[SNR < SNR_0] \quad (5)$$

Assuming the mobility scenario we have defined, it is possible to determine the value of the outage capacity for a given SNR_0 using Monte Carlo (MC) method. This method is proceeded according to the transmitter position distribution (which has been considered Gaussian and inside the room). For each point, the SNR is computed from (2) and from H expressions (3) or (4) depending on the configuration (LOS or diffuse). The outage probability is equal to the fraction of points whose SNR is below SNR_0 among the total number of points.

Figs. 2 and 3 present outage probability P_{out} versus SNR_0, estimated with MC method for LOS and diffuse configurations. The results have been plotted for different rates R_b.

As expected, we can see that, for both configurations, outage probability increases when the threshold value SNR_0 increases. Moreover this performance degradation also depends on the data rate of the transmission and becomes all the more significant as R_b increases. Even if the same behavior can be observed in LOS and diffuse configurations, the outage probability in the LOS configuration is more sensible to SNR_0.

For example, in the LOS configuration, if the system requires a SNR_0 of 15.6dB (to ensure a BER below 10^{-9} when the system is not in outage), and if the targeted outage probability is 10^{-3}, results reported in Fig. 2 show that the data rate has to be chosen below 3 Mbps. In diffuse configuration, if we consider the same outage probability of 10^{-3}, we observe in Fig. 3 that the rate has to be lower than 600Kbps.

In order to illustrate the performance in both configurations, we now determine the probability density function of the SNR inside the room by deriving P_{out} with respect to SNR. SNR distributions are presented in Fig. 4 for both configurations and for the same data rate of 1Mbps.

We can note that the SNR distribution in LOS configuration admits a minimum value $SNR_{min} = 17$dB corresponding to a case where the transmitter is placed on the floor, in a corner ($d = d_{max}$.) Maximum SNR value ($SNR_{max} = 35$dB) is obtained when the transmitter is placed beneath the receiver at a maximum height of 1.5m (due to mobility constraints). In diffuse configuration, the SNR distribution admits a greater spreading than in the LOS case. Moreover, the diffuse case presents a lower average SNR value ($\overline{SNR} = 27.44$dB) compared to LOS one ($\overline{SNR} = 29.5$dB).

These SNR distributions represent the non-stationnarity due to the particular mobility scenario we study for the indoor optical wireless channel in both LOS and diffuse configurations.

V. Outage Capacity

In this paper, we consider a binary input (due to the OOK modulation) and continuous output AWGN channel (due to the noise present over the optical channel).

For stationary channel, the capacity of binary input continuous output AWGN channel does not admit a close form. Thus, this capacity has to be evaluated by using [14]:

$$C(SNR) = \sup_{p(x)} \int_{-\infty}^{\infty} \sum_{i=0}^{1} p(y|x_i)p(x_i) \log \left(\frac{p(y|x_i)}{\sum_{k=1}^{n} p(y|x_k)p(x_k)} \right) \mathrm{d}y \quad (6)$$

where $p(y|x)$ are the conditional probabilities of the received signal and follow Gaussian distributions $\mathcal{N}(RHX, R_bN_0)$. $p(x)$ corresponds to the probability of the symbol x. Since the channel is symmetric (6) is maximized when $p(x = 0)$ and

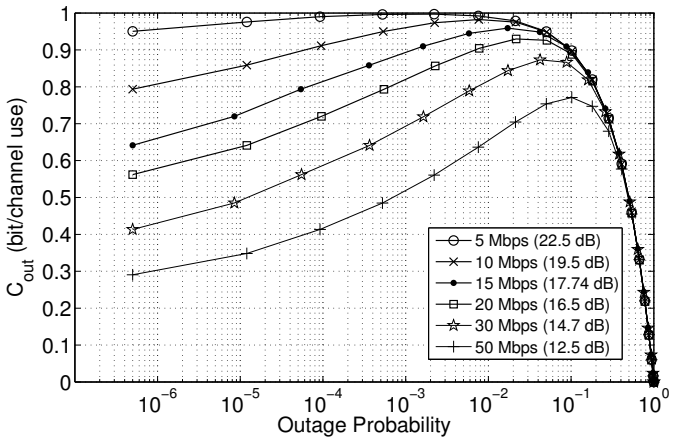


Fig. 5. Capacity of LOS wireless indoor channel for OOK modulation

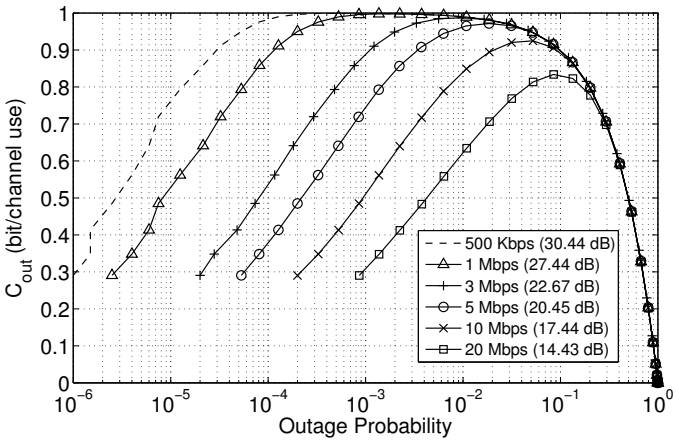


Fig. 6. Capacity of diffuse wireless indoor channel for OOK modulation

$p(x = 1)$ are equal to 0.5. The capacity is bounded between 0 and 1 due to the binary input.

For non-stationary (flat fading) channel, the capacity depends on the information available at the receiver [11]. In this paper, we assume that the receiver has full and perfect knowledge of the Channel State Information (CSI). This can be obtained by inserting pilot symbols during the transmission. At the receiver, these pilot symbols are used to evaluate instantaneous SNR (or equivalently, instantaneous H).

The outage capacity, which well describes the performance of quasi-static channel, is defined as the average information rate that can be received with a given outage probability which increases the information rate. The transmitter fixes a rate *a priori* and sends data over the channel of capacity $C(SNR)$ (see eq. (6)). With a given outage probability, the average information rate correctly received is [11]:

$$C_{out} = (1 - P_{out}(SNR_0))\, C(SNR_0) \tag{7}$$

Note that C_{out} is proportional to $(1-P_{out})$ which corresponds to the absence of transmitted information (*i.e.* a null capacity) during outage events.

Figs. 5 and 6 present outage capacity versus outage probability for LOS and diffuse configurations. In both cases, outage capacity is computed with (6) and (7) for different data rates corresponding to different $\overline{SNR}$ reported in parenthesis on the figures. The outage probability has been estimated using previously described MC method.

Outage capacity for both configurations is bounded between 0 and 1 due to the input constraint.

When P_{out} tends to 1, the receiver is always in outage and the maximum information rate that can be transmitted between transmitter and receiver tends to 0 ($C_{out} = 0$). On the other side, when P_{out} tends to zero, C_{out} attempts a minimal value equal to $C(SNR_{min})$ where SNR_{min} is the lower SNR that can be received in the room.

Between these two values of P_{out}, there is a given value of the outage probability maximizing the channel capacity. This value depends on the data rate *i.e.* $\overline{SNR}$, and increases when data rate or $\overline{SNR}$ decrease.

For LOS configuration, the maximal capacity is obtained for P_{out} belonging in $[10^{-3}, 10^{-1}]$ for data rates between 5 and 50Mbps. Data rates of 5Mbps and 50Mbps corresponds to $\overline{SNR}$ in LOS configuration of 22.5 and 12.5dB. The outage capacity is reduced when $\overline{SNR}$ decreases. For example, the maximal capacity varies from 1 to 0.75 bit/channel use for respectively 5Mbps and 50Mbps. Capacity equal to 1 means that there is no need to use FEC to achieve the maximal information rate. An outage capacity below 1 (*e.g.* $C_{out} = 0.75$) means that using error correction codes (of rate 0.75 in this example), a reliable transmission over the channel can be achieved.

In diffuse configuration same remarks can be done. The outage probability which maximizes the capacity is obtained in the interval $[10^{-3}, 10^{-1}]$ for gross data rate between 500Kbps and 20Mbps (*i.e.* $\overline{SNR}$ of 30.44dB and 14.43dB). The corresponding outage capacity is included in the interval $[1, 0.85]$. To compare the two configurations, we consider a gross data rate of 20Mbps. We can note in Figs. 5 and 6 that the maximal capacity is obtained for P_{out} of 10^{-2} in LOS configuration and is equal to 0.9 bit/channel use whereas it is of 0.8 in diffuse one for P_{out} around 10^{-1}. Consequently, diffuse channel presents lower capacities with higher outage probabilities than the LOS channel.

In the following, we illustrate FEC performance for a given LOS configuration.

VI. Performance Evaluation of LDPC codes

Different LDPC codes will be applied to the LOS channel in order to estimate the gain provided by FEC. A LDPC(N,K) is a linear block code and can be defined by its rate $r = K/N$, where K is the length of information bits and N is the codeword length [12]. The net data rate is thus reduced and can be expressed by $R_a = (K/N) \times R_b$, where R_b is the gross data rate of the transmission. We consider here regular LDPC codes wih a decoding process based on message passing

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

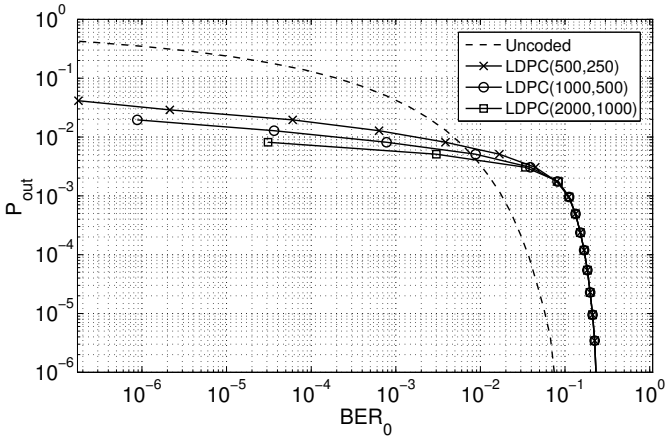


Fig. 7. Performance of LDPC coded transmissions in LOS configuration for OOK modulation and $R_b = 50$ Mbps

algorithm [12]. In order to improve the performance, the decoder uses soft-demodulation followed by the sum-product algorithm [13].

In order to estimate the performance of coded transmission for the two configurations, we introduce a threshold bit error rate BER_0. Outage probability can be thus expressed as a function of BER_0:

$$P_{out} = \Pr[SNR < SNR_0] = \Pr[BER > BER_0] \quad (8)$$

where $BER = f(SNR)$ is a monotonically decreasing function depending on the error correcting codes and the modulation scheme used. Since close form of this function is only known for uncoded transmission, the estimation of $f(.)$ function for coded transmission over stationary optical channel with OOK modulation and using regular LDPC codes has been done by simulation. We consider a gross data rate of 50Mbps and an outage probability of 10^{-3}. We can see in Fig. 5, that the LOS channel provides a capacity of 0.5 bit/channel use for this example. We have thus considered LDPC codes of rate 1/2 and of lengths 500, 1000 and 2000.

Fig. 7 presents the performance of coded and uncoded transmissions on the graph P_{out} versus BER_0 for LOS configuration.

As expected, we can see that LDPC of rate 1/2 can achieve lower P_{out} than in the uncoded case. Actually, the coded performances approach the capacity bound of 10^{-3}. Moreover, we can note that the outage probability decreases when the length of the code increases. For example, for a BER of 10^{-6}, the outage probability varies from 3×10^{-2} to 1×10^{-2} for respectively $N = 500$ and $N = 1000$, whereas the outage probability in the uncoded case is around 0.5.

Thus, LDPC codes provide an efficient way to improve the performance of the optical link and permit reducing the outage probability for a given transmission.

VII. Conclusion

In this paper, we have evaluated the performance of LOS and diffuse wireless optical transmissions in indoor environment by considering the mobility of the optical transmitter. For OOK modulation, we have estimated the outage probability of each link configuration using simulations. From the results, one can determine the attainable data rate for a given quality of service. We have then studied the system potentiality by evaluating the outage capacity. In each configuration case, the outage probability that maximizes the capacity for a given data rate has been obtained. We have observed that the diffuse channel presents lower capacities with higher outage probabilities than the LOS one.

The performance gain for LOS case with regular LDPC codes of different lengths has been finally presented. The results have shown that this constitutes an interesting solution to enhance the performance (in terms of outage probability) of a mobile indoor IR transmission.

References

[1] J. M. Kahn and J. R. Barry, "Wireless infrared communications," *Proceedings of the IEEE*, vol. 85, no. 2, pp. 265-298, 1997.

[2] F. R. Gfeller and U. Bapst, "Wireless in-house data communication via diffuse infrared radiation," *Proceedings of the IEEE*, vol. 67, no. 11, pp. 1474-1486, 1979.

[3] H. Park and J. R. Barry, "Modulation analysis for wireless infrared communications," in *ICC ?95 Seattle, Gateway to Globalization, IEEE International Conference on Communications*, vol. 2, pp. 1182-1186, 1995.

[4] J. M. Kahn, W. J. Krause, and J. B. Carruthers, "Experimental characterization of non-directed indoor infrared channels," *IEEE Transactions on Communications*, vol. 43, no. 234, pp. 1613-1623, Apr. 1995.

[5] J. Anguita, I. Djordjevic, M. Neifeld, and B. Vasic, "Shannon capacities and error-correction codes for optical atmospheric turbulent channels," *Journal of Optical Networking*, vol. 4, no. 9, pp. 586-601, 2005.

[6] J. Li and M. Uysal, "Optical wireless communications: system model, capacity and coding," in *Proceedings of IEEE Vehicular Technology Conference*, *VTC 2003-Fall*, vol. 1, pp. 168-172, 2003.

[7] A. A. Farid and S. Hranilovic, "Outage Capacity Optimization for Free-Space Optical Links With Pointing Errors," Journal of Lightwave Technology, vol. 25, no. 7, pp. 1702-1710, Jul. 2007.

[8] A. Belmonte and J. M. Kahn, "Fundamental limits of diversity coherent reception on atmospheric optical channels," in *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, pp. 1621-1625, 2009.

[9] S. Hranilovic, "On the Design of Bandwidth Efficient Signalling for Indoor Wireless Optical Channels," *International Journal of Communication Systems - Special Issue on Indoor Optical Wireless Communication Systems and Networks*, Wiley Interscience, vol. 18, no. 3, pp.205-228, April 2005.

[10] R. Ramirez-Iniguez and R. J. Green,"Indoor optical wireless communications," *IEE Colloquim on Optical Wireless Communications*, pp. 14/1-14/7, 1999.

[11] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[12] S. J. Johnson, *Iterative Error Correction: Turbo, Low-Density Parity-Check and Repeat-Accumulate Codes*, 1st ed. Cambridge University Press, 2009.

[13] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*, 1st ed. Cambridge University Press, 2003.

[14] J. Proakis, *Digital Communications*, 4th ed. McGraw-Hill Science/Engineering/Math, 2000.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

A Resource Management Architecture for Mobile Satellite-based Communication Systems

Philipp Drieß, Florian Evers, Markus Brückner
Integrated Communication Systems Group
Ilmenau University of Technology
Helmholtzplatz 5
98693 Ilmenau, Germany
Email: {philipp.driess, florian.evers, markus.brueckner}@tu-ilmenau.de

***Abstract*—Today's Quality-of-Service (QoS) architectures for packet switched networks, especially reservation-based architectures such as IntServ, strongly depend on transmission systems with link characteristics that do not change over time. In wireless network access technologies however, this requirement makes implementation of QoS challenging as links change due to effects such as shadowing or fading caused by node mobility. In this paper, a novel cross-layer reservation-based QoS system is presented. It is able to deal with changing link conditions and notifies the affected applications with the help of feedback messages. This allows graceful degradation in compliance with the requirements of the applications. The architecture presented in this paper focuses on a satellite-based network for rescue teams during a disaster scenario. Based on a geostationary satellite and highly mobile ground stations, this network is characterized by long transmission delays and small, changing link capacities. The system offers a resource reservation scheme on the physical layer which is integrated with the approach presented here to design a cross-layer resource management architecture. This results in a reservation system for high-latency, low-capacity and unstable links.**

***Index Terms*—quality of service; satellite communication; mobile communication; IntServ; signaling**

I. Introduction

In contrast to wire-based transmission systems such as Ethernet, the transmission conditions of wireless systems are considered *unstable*. This is especially true if nodes become mobile: laptops that are carried around experience different kinds of fading effects, and mobile satellite terminals are constantly affected by trees, clouds and other obstacles that impair the line-of-sight transmission to the satellite.

Such links with changing conditions can not be avoided, which leads to problems regarding support for Quality-of-Service (QoS). Reservation-based schemes like "Integrated Services" (IntServ [1]) depend on networks with *stable* links for their capacity management, which fails if the available capacity is a dynamic parameter. In contrast, "Differentiated Services" (DiffServ [2]), an architecture that is based on the differentation of traffic into classes with specific properties, does not offer reservations at all and thus is not able to offer guarantees to the applications.

Depending on the intended use case having guarantees might be a requirement. In the research project "Mobile Satellite Communications in Ka-Band" (MoSaKa, see [3] for an introduction), a satellite-based communication system is developed to support rescue teams in disaster scenarios. In such a system, voice communication is one very important application, ideally in combination with video. As satellite resources are scarce, not all communication attempts can be admitted. However, continuous communication streams like voice conversations are not the only traffic in disaster communication systems. They are also used for all kinds of data traffic like digital maps, status reports, and position information. Therefore, a packet-switched approach based on a protocol suite like TCP/IP is the most flexible approach to build such a network. To serve important applications like voice reliably, a reservation-based scheme that assures QoS can be implemented, causing the aforementioned problem regarding the unstable characteristics of the link to the satellite. The availability of a QoS system coping with those limitations will be a key factor for being able to use packet-based satellite communication systems as backbones, especially in disaster scenarios.

In this paper the MoSaKa QoS system, a novel reservation-based QoS architecture that is able to cope with unstable links, is presented. The focus is on satellite-based networks. This empowers rescue teams to communicate in environments without communication infrastructure, which is the research area the MoSaKa project is looking at. However, the algorithms shown in this paper could also be applied to other QoS-enabled wireless transmission systems, such as IEEE 802.11e.

The remainder of this paper is organized as follows: Section II describes the environment for which the presented solution was designed. From this the system requirements are derived in Section III. Section IV gives an overview over the related work. Following this Sections V and VI present the proposed architecture and the test environment which is being built to evaluate the approach. The paper finishes with an outlook to future work and a conclusion.

II. The MoSaKa Research Project

The MoSaKa project funding the research presented in this paper aims at developing a complete satellite communication stack from the antennas up to the QoS management, including antenna tracking systems and a decentralized resource allocation scheme. In this paper, the main focus is on the QoS system as it is seen by the higher layers. Layer 2 and below are

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

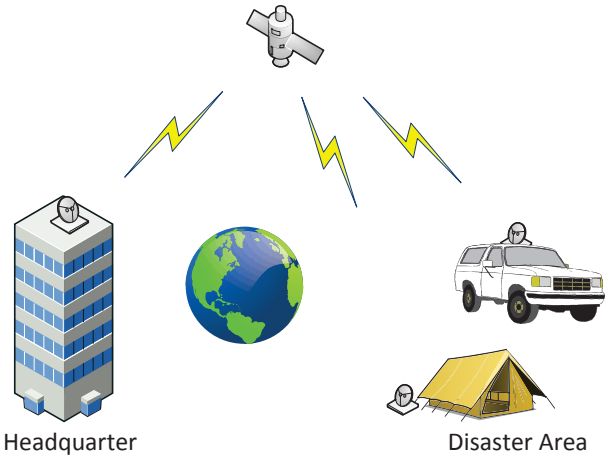


Fig. 1. Typical use cases for MoSaKa entities: fixed, nomadic and mobile terminals

only mentioned insofar as they are required to explain design decisions taken in the higher layers.

Figure 1 shows a typical usage scenario: some nomadic and mobile terminals deployed in a remote area use the satellite link to communicate with their headquarter. Each terminal uses a dynamic set of services resulting in individual traffic demands. The communication link from each terminal to the satellite is considered unstable: the link quality fluctuates with the movement of the terminal and changing environment conditions. Most of these fluctuations are short, but some may persist for a longer time span.

Large-scale disaster recovery operations cause a huge demand for communication. Satellite links are a – comparatively – scarce resource with only a small capacity and long delays (for geostationary orbits at a height of $\approx 36\,000$ km the time-of-flight is already more than 100 ms for one direction). These two effects have to be considered if a system-wide QoS infrastructure is implemented.

III. Requirements

Based on the scenario given in Section II, a set of requirements that a QoS infrastructure has to fulfill can be derived.

Efficient handshakes

The main issue in designing an efficient signalization scheme for QoS requirements is the long transmission delay introduced by the satellite link. With a round trip time of ≈ 400 ms, complex handshakes with multiple messages travelling back and forth are imposing an unacceptable overall latency.

Efficient handling of link instability

Today's QoS systems assume stable links with static resources that can be utilized for reservations. This assumption is no longer valid in mobile, satellite-based communication systems or even in mobile communication systems in general. Over the time, the propagation conditions are subject to change. The QoS infrastructure presented in this paper must be able to cope with unstable link conditions.

Cross-layer link usage optimization

Satellite-based communication with multiple terminals takes place on a shared broadcast medium. To enable parallel transmissions via one single satellite the MoSaKa physical and MAC layers have to distribute the available link spectrum to all terminals that compete for resources. This happens with respect to the individual resource demands of each terminal. Beforehand these resource demands have to be derived from higher-layer QoS requirements that originate from the applications.

Due to the long delay of the broadcast medium the resource assignment procedure takes place in a distributed manner without central coordination and without any point-to-point negotiation. If the link share decreases the higher layer reservations may not fit anymore. In that case, the QoS system has to evaluate all admitted reservations based on their properties to keep as many of them as possible active.

A resource management system suitable for mobile satellite communication has to address those requirements. Existing solutions fall short in one or the other aspect prompting the development of a new architecture for the MoSaKa project.

IV. Related work

QoS architectures such as IntServ [1][4][5] or DiffServ [2][6] are well known and have a wide range of acceptance. Nevertheless, they have a variety of issues regarding unstable link conditions.

IntServ

IntServ is an architecture that offers hard guarantees regarding QoS parameters. Applications request reservations via a signaling protocol such as the "Resource Reservation Protocol" (RSVP [7]) or "Next Steps in Signaling" (NSIS [8][9][10]) to announce their individual traffic requirements. On each node along the transmission path an IntServ entity manages and monitors the traffic regarding the requested resources.

Applying IntServ upon an unstable link leads to problems if the link capacity starts decreasing. This results in a situation where the sum of all accepted reservations does not fit into the link budget anymore and the reservations are violated. As no feedback mechanism is available, the system has to withdraw reservations. Affected applications can only deal with this situation by reserving a new path with different parameters or ceasing communication altogether. Signaling new paths causes additional message load on the already limited link, contributing further to the congestion.

DiffServ

One of the problems of IntServ in large-scale networks is its bad scalability. Due to the state kept in every intermediate node, IntServ installations do not scale to Internet-sized networks. This prompted the development of DiffServ which is based on differentiation of traffic into classes which can be treated differently by the network. This allows the assignment of transmission priorities to distinguish different types of traffic.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

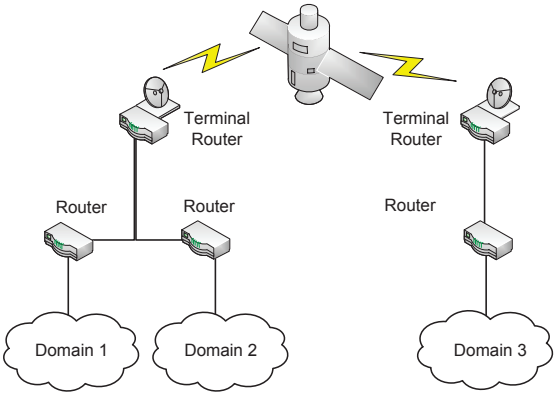


Fig. 2. The network of a scenario where the MoSaKa QoS architecture is deployed

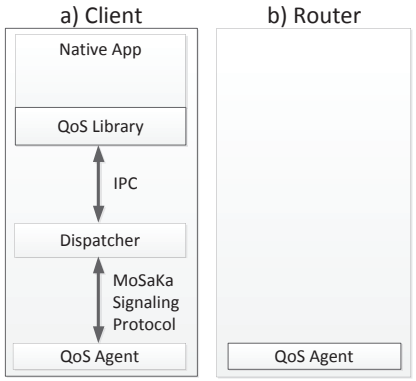


Fig. 3. Two kinds of nodes exists in the MoSaKa network: clients and routers.

As DiffServ does not support the reservation of a communication path, it does not offer any guarantees. Excessive traffic in a single class might exceed the link capacity, causing packet loss for all affected applications.

For use in mobile satellite environments, both approaches lack certain desirable features. This prompted the development of a new approach based on IntServ. While the scalability of IntServ to large networks might be a problem, this will not be an issue in the system at hand. However, the guaranteed reservation of communication paths as offered by a reservation-based system is crucial in a disaster scenario. MoSaKa aims at solving the challenges arising from unstable links while keeping the reservation features of IntServ.

V. The QoS architecture of MoSaKa

One part of the MoSaKa project has the goal to build a reservation-based QoS architecture that is able to cope with unstable link conditions. Therefore, an IntServ-like approach was chosen, which introduces management entities on each intermediate node as well as on each end system. These entities are aware of all reservations that pass the respective node. In the depicted scenario, static routing in the backhaul is assumed, which ensures that each packet of a flow always takes the same route through the network.

The resulting topology is depicted in Figure 2. The central component is the satellite-based communication system with a geostationary satellite and multiple terminals as ground stations. The satellite link is considered to be a bottleneck with high latency. The terminals act as routers for the IP protocol, and attach local networks to the satellite network.

The result is a QoS architecture without a central coordinator. Unfortunately, this approach inherits two issues of IntServ: it has scalability problems and might fail if the links are unstable. The former can be neglected with the depicted use case in mind, but the latter will be discussed in this paper.

A. Software components

The components introduced by the MoSaKa QoS architecture are depicted in Figure 3. There are two main components: the QoS Agent and the Dispatcher.

1) The QoS Agent: The QoS Agent is a management entity that exists on each node of the network including routers and clients. This entity is aware of all ongoing reservations that pass the node and has an overview of the transmission resources of each interface that the node possesses. This allows the QoS Agent to decide whether a subsequent reservation can be admitted or has to be rejected. For the purpose of transmitting reservation request, a signaling protocol such as RSVP or NSIS is required. The QoS Agent intercepts protocol messages and interprets them as necessary. On the satellite terminal, it also communicates with the lower layers to detect if the link deteriorates.

QoS components like traffic metering and shaping are highly dependent on the underlying operating system of a node. It is the task of the QoS Agent to adapt the high-level reservations to the QoS primitives available on the host to allow a wide deployment of the architecture in heterogeneous networks. Each agent, therefore, consists of a generic part handling the signaling and admission control and a system-specific part configuring the underlying operating system services.

2) The Dispatcher: The Dispatcher is an optional component that is only required if a given node has applications running on it, making it a client. A dispatcher acts as a broker between the applications running on the client and the QoS system in the network. The applications talk to the Dispatcher using “interprocess communication” (IPC). The Dispatcher handles all QoS-related interaction with the network relieving the application from doing so. Additionally it serves as an entry point for requests and notifications from the network, decoupling the local application structure and the state saved along the communication path. From the network point of view the Dispatcher is the entity that holds a reservation and renews it as necessary.

Reservations are always triggered by an application. The Dispatcher merely acts as a proxy. Therefore applications are a part of the MoSaKa QoS architecture as well and need to be modified to take full advantage of the system. One has to distinguish QoS-enabled applications, legacy applications and

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

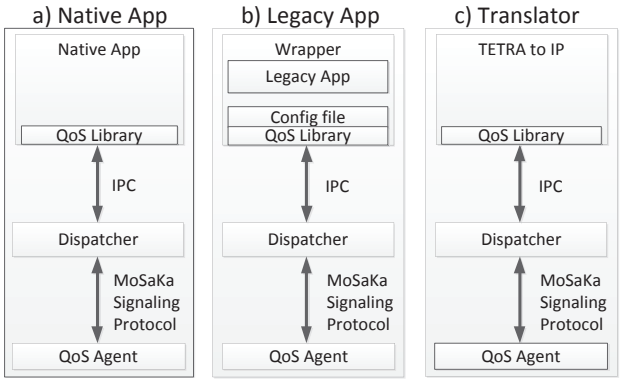


Fig. 4. Three kinds of applications are distinguished: Native, legacy and translator applications.

translator applications.

QoS-enabled applications: As shown in Figure 4 a, a QoS-enabled architecture includes the MoSaKa QoS library. This library offers an high-level API to interact with the QoS architecture and allows the programmer to easily request transmission resources or to be notified if an active reservation fails or deteriorates. Such an application is aware of its traffic demands and is able to request the appropriate amount of resources before it starts transmitting. Additionally, it is aware of the fact that feedback messages may arrive that indicate that the reservation is affected by the current transmission conditions.

Legacy applications: All IP-based applications that exist today are considered as legacy applications. They are not aware of an API to request transmission resources, resulting in traffic that is not known to the local QoS Agent. Two approaches are possible: this traffic can be considered as "best effort" traffic, which may or may not pass a bottleneck in the network, or it can be reserved with the help of a "wrapper application" (Figure 4 b).

Such a wrapper loads a predefined set of QoS requirements from a configuration file, initiates the reservation process and then, if successful, executes the legacy application. In that case, the application does not need to know anything about the QoS system, but benefits from it nevertheless, as the required resources are reserved.

The reservation is held all the time even if the application does not emit traffic. Even worse, to initiate a reservation, the endpoint must be named, which limits the application to a given set of predefined peers. However, such a wrapper can be seen as an intermediate solution until the affected applications implement the QoS scheme.

Translator applications: As a third kind of applications, a translator application, acts as a gateway to other kinds of reservation schemes or networks such as circuit-switched telephony systems. Such an entity is a special case of a QoS-enabled application (Figure 4 c).

In disaster scenarios, connections with other network types such as "Terrestrial Trunked Radio" (TETRA [11]) may be required. A dedicated gateway node with a TETRA base station and a translator application installed on it can interconnect both networks, allowing TETRA terminals to make telephone calls to the headquarter via the satellite. The translator application is aware of the required resources of a TETRA channel, as the traffic requirements of the codecs involved are known. This allows it to send suitable reservation requests into the MoSaKa network.

B. The QoS-enabled network

The network consists of two kinds of nodes: intermediate nodes are referred to as routers, and end systems are referred to as clients.

As routers have no applications running on them, the only entity required here is the QoS Agent. The routers that are connected to the satellite system are referred to as terminals. On such a terminal the QoS Agent is equipped with additional capabilities to manage the link to the satellite.

Clients, as they are considered as user equipment, have applications running on them additionally requiring the Dispatcher as a bridge to the network.

On each network node the QoS Agent has to configure the local packet forwarding entity of the operating system to stop misbehaving applications from congesting the outgoing interfaces. It should be impossible that traffic, that exceeds the capacity of the outgoing link, causes packet loss for flows that have been negotiated before. This is achieved relying on platform specific mechanisms to control traffic flow like "Traffic Control" (tc) and "Netfilter" on Linux or the MoSaKa MAC scheduler on the satellite terminal.

The signaling scheme

On each client QoS-enabled applications communicate with the local Dispatcher via an API offered by the MoSaKa QoS library. Through this API the application informs the QoS System about the amount of resources it requires for a transmission to a well-defined peer. The Dispatcher creates a reservation request signaling message that it sends to its peer entity, the Dispatcher on the destination node. All signaling messages are intercepted by each QoS Agent along the path to the destination, allowing them to decide whether to accept or to deny the reservation request. If such a reservation has to be denied because of insufficient remaining link resources, a negative acknowledgement is sent back to the initiating Dispatcher. This results in a deletion of the pending reservation on all intermediate nodes and leads to a negative acknowledgement to the application via the QoS Library.

In case of success, the QoS Agent on the destination node informs the local Dispatcher of an incoming reservation request. This Dispatcher may be aware of local applications as they are allowed to register to it beforehand. Nevertheless, it sends an acknowledgement back to the originator. This message is intercepted by all QoS Agents again and results in an orderly created reservation along the whole path.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

C. Feedback mechanism

To deal with changing link conditions, MoSaKa adds a feedback mechanism to the signaling protocol. Such feedback messages originate from an QoS Agent observing a deteriorating link on its outgoing interfaces and are sent to all applications that hold reservations affected by this degradation.

On each node, the link hardware is monitored by the local QoS Agent. Additionally, this entity is aware of all active reservations that involve this link, allowing it to decide whether the remaining capacity is still high enough to serve all reservations. If the capacity falls below the amount of reservations, the QoS Agent starts to optimize. Optimization is done by building a set of allowed reservations starting from the one with the highest priority. Reservations are incrementally added to the set if there is still capacity available. This simple optimization algorithm is suited well for highly hierachical communication environments like desaster recovery operations.

After optimization the QoS Agent has a list of reservations that still have enough resources and a list of reservations that do not fit into the link anymore. Instead of cancelling these reservations as architectures such as IntServ would have to do, the QoS Agent of MoSaKa is able to put affected reservations "on hold". Such a reservation is still known to the whole path, but can not be utilized for the moment. If the link recovers shortly later, the reservation is reactivated by another feedback message. As the MoSaKa scenario states that link degradations are short in nature, this approach allows a reservation scheme with a low amount of signaling messages. Applications do not need to actively poll the network for free resources which would introduce a high signaling load. Nevertheless, if the link stays degraded for a longer time span, the QoS Agent may cancel the reservation to prevent congesting the network with reservations that cannot be served anyway.

In the MoSaKa QoS architecture, signaling messages are usually exchanged between Dispatchers on two peer nodes, and are intercepted by all QoS Agents on each intermediate node including the end nodes that run the Dispatchers. If a given reservation has to be suspended, the QoS Agent creates signaling messages and sends them to both Dispatchers, allowing all other QoS Agents to notice that this reservation is currently "on hold".

If a signaling message arrives at a Dispatcher, it relays it to the respective application which triggers a trap in the QoS library informing the application about an accepted, suspended, resumed or cancelled reservation.

This feedback scheme is new and allows a graceful degradation of communication. To underline this, one of the most important applications of the MoSaKa scenario is analyzed: Video chat.

D. Impact on Video chat

If a user starts a video chat session with the headquarter, the video chat application tries to reserve resources for the video data and for the audio data separately. As a video chat session is bidirectional, the reservation requests resources for both directions at the same time, allowing the signaling handshake to complete fully after just one round trip. If the reservation handshake is completed without rejections from intermediate systems the path is active and can be utilized.

If the satellite link deteriorates, this is noticed by the QoS Agents on the satellite terminals. They start the optimization process which results in the less important video streams to be put "on hold" to keep the audio streams active. A feedback message is sent to the Dispatchers at both ends of the path, resulting in the deactivation of the video stream in the application. If the link recovers, another feedback allows the video stream to be resumed. Applications may provide a visual indication based on the network state to make the proess transparent and increase user satisfaction. If the link fails to recover the reservation is cancelled by the network. This frees resources permanently for reuse by other applications.

The "on hold" state allows the system to bridge short link degradations that are common in mobile satellite communication without reconfiguring the whole path causing large signaling effort. The feedback mechanism allows applications to intelligently react to those changes in the network. Especially for satellite links with long delays and a low capacity, such a scheme is essential to operate as desired.

E. MoSaKa Satellite Terminals

To check whether all active reservations fit into the current link capacity, the QoS Agent has to obtain this information. For that purpose, technology-dependent functionality is required to interact with Ethernet, IEEE 802.11e or the MAC- and PHY layers of the MoSaKa satellite terminals.

The MoSaKa satellite link offers QoS-enabled lower layers. From the physical layer point of view the satellite link is always a shared medium. Each terminal can be received (although not necessarily decoded due to signal quality issues) by every other terminal via the satellite. Therefore it is necessary to allocate parts of the link spectrum to specific sender terminals to prevent collisions. To accommodate for changing link conditions this allocation is not static but takes place every $250\ \mathrm{ms}$. Each active terminal is assigned a short time slot on the lower layer (L2) signaling channel in this period and broadcasts its resource request to all other terminals. Based on this information, each terminal applies the same resource assignment procedure and comes up with the same resource allocation vector for the next $250\ \mathrm{ms}$ data transmission frame. A reservation on the lower layers is valid for only one slot, and has to be renewed continuously by using the L2 signaling channel.

If the link between one terminal and the satellite deteriorates, the QoS Agent on this terminal gets informed about a lower amount of transmission resources that this terminal got assigned for the next data transmission frame. This allows the QoS Agent to check whether high-level reservations and available link share still match and take appropriate actions if they do not.

VI. Evaluation

The presented architecture will be implemented as a proof-of-concept for the Linux operating system. To evaluate the

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

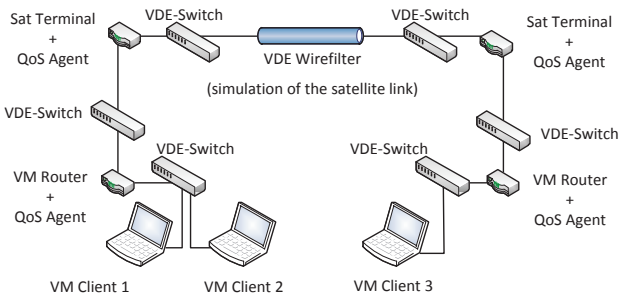


Fig. 5. The MoSaKa virtual testbed based on QEMU/KVM and VDE

implementation a virtual network environment based on QEMU/KVM [12] is currently set up. Different Linux-based guest systems are interconnected using "Virtual Distributed Ethernet" (VDE [13], by Virtual Square). Heart of the testbed is the satellite link emulated by wirefilter, a part of VDE that allows to set up virtual Ethernet connections with various parameters such as the data rate, delay or the packet loss rate. At each end of the emulated link resides a Linux-based guest system acting as a the MoSaKa satellite terminal. Connected to them are networks of different structure depending on the scenario. A possible scenario is shown in Figure 5. Each part of the network contains an additional QoS Agent as well as a set of clients acting as signaling and traffic sources.

In parallel to this virtual testbed designed to test the high-level parts of MoSaKa, the complete stack including the terminal hardware is currently being implemented. This "real world" setup will allow tests of the whole communication path developed during the project. The resulting architecture will include all layers, from the QoS-enabled applications down to the antenna. The system is based on a virtual satellite placed on a tower. A motion emulator and a channel simulator provide the infrastructure to simulate the influence of mobility on the Ka-band channel. A directional antenna on a 3D rotor acts as the mobile terminal. The motion emulator is driven by data from measurement campaigns carried out earlier to create a communication environment as it is expected "in the wild". Using this realistic emulation environment close-to-real measurement results on all layers are expected from this implementation.

VII. Conclusion and future work

This paper presented a novel QoS architecture called "MoSaKa QoS" for mobile satellite communication links. Existing approaches like IntServ are not suited for such an environment with unstable, long delay links. The optimistic, bi-directional signaling architecture of MoSaKa QoS with the included feedback mechanism supports reaction to link changes while minimizing the number of messages exchanged. Cross-layer resource optimization spanning layers 1 to 3 enables a better usage of the currently available channel, maximizing the user experience.

Future research should investigate into the possibilities opened up by the MoSaKa feedback mechanism. Modern audio and video codecs offer various output profiles with different data rates and quality settings. An integration with the QoS system might provide further options for graceful degradation of the link.

From the architecture point of view further research might look into alternative QoS models based on probability distributions instead of hard thresholds. Equally interesting are novel reservation models which provide more powerful ways to express requirements, enabling the system to adapt better to changing conditions without consulting the application. Another open question is the extension of the optimization algorithm to more general use cases without a strong hierarchy among communication paths.

At the moment MoSaKa uses a static IPv6 routing setup. In the future the system should adapt to various routing protocols to enable network-level node mobility.

One final research direction is the integration of legacy applications into the system. Applications that cannot be adapted to the MoSaKa system could be integrated using translator applications/application-level proxies or DiffServ-like classification approaches.

References

[1] J. Wroclawski, "The Use of RSVP with IETF Integrated Services," *RFC 2210*, September 1997.

[2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," *RFC 2475*, December 1998.

[3] M. Hein, A. Kraus, R. Stephan, C. Volmer, A. Heuberger, E. Eberlein, C. Keip, M. Mehnert, A. Mitschele-Thiel, P. Driess, and T. Volkert, "Perspectives for Mobile Satellite Communications in Ka-Band (MoSaKa)," in *EuCAP 2010: The 4th European Conference on Antennas and Propagation*, Barcelona, Spain, 04 2010.

[4] J. Wroclawski, "Specification of the Controlled-Load Network Element Service," *RFC 2211*, September 1997.

[5] S. Shenker, C. Partridge, and R. Guerin, "Specification of Guaranteed Quality of Services," *RFC 2212*, September 1997.

[6] K. Nichols, S. Blake, F. Baker, and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," *RFC 2474*, December 1998.

[7] A. Mankin, Ed., F. Baker, B. Braden, S. Bradner, M. O'Dell, A. Romanow, A. Weinrib, and L. Zhang, "Resource ReSerVation Protocol (RSVP) – Version 1 Applicability Statement Some Guidelines on Deployment," *RFC 2208*, September 1997.

[8] R. Hancock, G. Karagiannis, J. Loughney, and S. V. den Bosch, "Next Steps in Signaling (NSIS): Framework," *RFC 4080*, June 2005.

[9] J. Manner, G. Karagiannis, and A. McDonald, "NSIS Signaling Layer Protocol (NSLP) for Quality-of-Service Signaling," *RFC 5974*, October 2010.

[10] E. G. Ash, E. A. Bader, E. C. Kappler, and E. D. Oran, "QSPEC Template for the Quality-of-Service NSIS Signaling Layer Protocol (NSLP)," *RFC 5975*, October 2005.

[11] E. Re, M. Ruggieri, and G. Guidotti, "Integration of TETRA with Satellite Networks: A Contribution to the IMT-A Vision," *Wirel. Pers. Commun.*, vol. 45, pp. 559–568, June 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1363306.1363328

[12] F. Bellard, "QEMU, a fast and portable dynamic translator," in *Proceedings of the annual conference on USENIX Annual Technical Conference*, ser. ATEC '05. Berkeley, CA, USA: USENIX Association, 2005, pp. 41–41. [Online]. Available: http://dl.acm.org/citation.cfm?id=1247360.1247401

[13] R. Davoli, "VDE: Virtual Distributed Ethernet," in *Proceedings of the First International Conference on Testbeds and Research Infrastructures for the DEvelopment of NeTworks and COMmunities*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 213–220. [Online]. Available: http://dl.acm.org/citation.cfm?id=1042447.1043718

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Traffic Evaluation of a Claim-based Single Sign-On System with Focus on Mobile Devices

Mateusz Khalil
Fraunhofer Fokus
Berlin, Germany
Mateusz.khalil@fokus.fraunhofer.de

Yacine Rebahi
Fraunhofer Fokus
Berlin, Germany
Yacine.rebahi@fokus.fraunhofer.de

Simon Hohberg
Fraunhofer Fokus
Berlin, Germany
Simon.hohberg@fokus.fraunhofer.de

Pascal Lorenz
University of Haute Alsace
Colmar, France
Lorenz@ieee.org

Abstract— **The work on Web services security is not that new; however, the provided solutions either are not efficient, or applicable only to some special cases. As an example, the standardization group *OASIS* specified a huge amount of standards, which are unofficially called *WS-**. Unfortunately, not all of these standards are implemented by modern frameworks and their applicability to mobile Web services is also questionable. As the WS-* technologies are very useful in realizing single sign on (SSO) solutions, we discuss, in this paper, an implementation prototype for a single sign-on system with WS-* standards and evaluate the resulting network traffic while focusing on mobile devices. To be more precise, two implementations (one based on NetBeans and another one developed by the authors themselves) were achieved and their corresponding results were compared. It appears that the authors' implementation performs better that the NetBeans one and reduces the traffic generated from 85% to 50%.**

Keywords-Web traffic; Single Sign-On; Mobile devices; SOAP; OASIS; WS-*.

I. INTRODUCTION

A crucial benefit for the emerging Web services' architectures is the ability to deliver integrated, interoperable and secure solutions. Ensuring the protection of Web services from attacks and misuse through the enforcement of comprehensive security models is critical.

The work on Web services security is not that new, however the provided solutions either are not efficient, or applicable only to some special cases. In the literature, there are some suggested security mechanisms, namely, Security Assertion Markup Language (SAML) [1], which is an XML-based standard for exchanging authentication and authorization data between security domains, that is, between an identity provider and a service provider. SAML was specified by the OASIS [2] Security Services Technical Committee and was intended to solve the *Web Browser Single Sign-On* (SSO) problem. Single sign-on solutions are abundant at the intranet level (using cookies, for example) but extending these solutions beyond the intranet has been problematic and has led to the proliferation of non-interoperable proprietary technologies.

The standardization group *OASIS* [2] specified a huge amount of standards, which are unofficially called *WS-**[3]. Unfortunately, not all of these standards are implemented by modern frameworks and their applicability to mobile Web services is also questionable. As the WS-* technologies are very useful in realizing single sign on (SSO) solutions, we discuss, in this paper, two implementation prototypes (one based on the NetBeans [16] technology and the other one not) for a single sign-on system with WS-* standards and evaluate the resulting network traffic while focusing on mobile devices.

The outline of the rest of this paper is as follows. Section two provides an overview of the techniques needed by the subsequent parts. Section three discusses the proposed approach and section, four presents the experimental results. Finally, section five concludes the paper.

II. BACKGROUND

A. Single Sign-On

Single sign-on (SSO) is an identity management model where the user needs to provide his credentials only once and will stay authenticated against the realm, which may include multiple services within the concerning *Circle of Trust*. There are many different implementations of SSO with prominent representatives like Kerberos [4] or OpenID [5]. Main parts of a SSO system in a decoupled claim-based scenario are Client (C), Relying Party (RP) providing a service and Identity Provider (IdP) in the role of a STS. Usually, the first step is a service request of C to RP. Then RP forwards the Client's request to a trusted IdP, because

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

RP requires identity information from C in order to check a defined authorization policy. To prove authorization C needs to provide a service token to RP. This service token is provided by the IdP when C authenticates using its credentials. Figure 2 shows a sequence how such a service token is requested. In the sequence in the figure, C first requests a security context token (SCT) sending a request security token (RST) message containing C's credentials. The STS (IdP) answers by sending a request security token response (RSTR) containing the SCT if C could be authenticated. Now C can use this valid SCT to request the service token from the STS (IdP) in the next step. Hereon, the STS validates the SCT and provides the service token and any requested attributes. With this service token and the attributes C is able to request the service (RP), so that the service token is validated successfully and the service policy is fulfilled. Finally, the service (RP) responds to the request.

The SCT's and the service token's integrity are commonly protected by a signature. In some solutions the token is not forwarded by C, but is delivered directly from STS (IdP) to the service (RP). Because there is no need for C to authenticate again after C has received a SCT from STS (IdP) until it has expired and C can use this SCT for further requests to STS (IdP), this mechanism is called single sign-on.

B. Related Work

WS-* [3] contains many specifications defined by the standardization group OASIS. The specifications deal with security issues as well as reliability, transactions and others. In this paper, we focused on the key security-related specifications *WS-Security*, *WS-Trust*, *WS-SecureConversation* and *WS-SecurityPolicy* [2]. These specifications describe how SOAP-messages [6] need to be built with the objective to maintain interoperability. In an SSO system: C, RP and IdP may be developed independently but can interact because of the defined languages specified in WS-*.

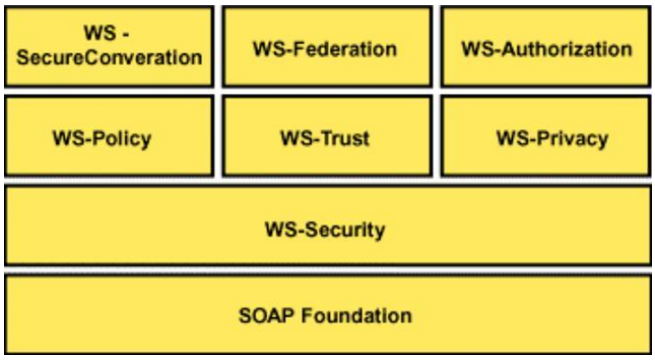


Figure 1: WS secure conversation

WS-Security defines how authentication and authentication information can be stored in the SOAP-Header. Basically other standards and specifications can be embedded into WS-Security.

WS-Trust is a specification which can be regarded as a Web service definition which may validate, create, renew or delete tokens. This Web service is commonly called Security Token Service (STS). The token is used as a proof of authentication (and may be limited in time). It may also contain authorization information.

SOAP is stateless which leads to inefficient usage of bandwidth if all security information needs to be transmitted in every SOAP-message. This problem is solved by **WS-SecureConversation** which provides methods for creating and protecting security sessions. Once a security context has been established, tokens, claims and key information can be stored within the concerning security context.

WS-SecurityPolicy describes a language for the definition of security requirements. For instance a valid security policy may require that the SOAP-body needs to be encrypted using a certain algorithm, expected token types and how the response message needs to be secured. The main benefit is that this policy can be published, which means that the consumer may adapt himself to the requirements. Another aspect of the policy driven development is its readability and efficiency concerning development effort.

Finally, the XML security recommendations consisting of **XML-Encryption** and **XML-Signature**, which were introduced by W3C [7], specify how to encrypt and sign any elements within an XML document. This implies that *XML-Encryption* and *XML-Signature* can be applied in SOAP-messages as well.

III. Our Approach

Web services have many benefits and can be used in various scenarios. Nowadays, most security dependent Web services are secured by SSL only. For instance we may have a production chain where many producers participate in. Assuming that parts of the SOAP message are confidential and have to be readable by one certain producer. SSL encrypts on transport level but an encryption on message level is required. This is where WS-* and XML security standards come into play. These technologies are also very useful in order to realize a single sign on (SSO) system. Especially in the production chain example, identity management of all participating parties has to be considered.

As not all of these standards are implemented by modern frameworks, a description of critical features provided in *Apache Axis2*, *Apache CXF* and *Glassfish Metro* is crucial.

A. Frameworks

Apache Axis 2 [8], Apache CXF [17] and Metro Glassfish [9] are popular Web service frameworks. They act as SOAP engines and have additional libraries which provide security functionalities. We have summarized which specifications

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

and features are supported by these frameworks. In our opinion *Metro Glassfish* in combination with *NetBeans* is the most effective choice (we refer to Table 1). NetBeans [16] is an extensible integrated development environment (IDE) providing necessary tools for developing desktop, enterprise, Web and mobile applications. Metro is simply an open source Web service stack that is a part of the Glassfish project.

B. Use Case

We created a simple Web service with *NetBeans* which was secured by WS-* technologies using *Metro Glassfish*. In particular WS-Security (UsernameToken Profile) WS-Trust, WS-SecureConversation, WS-SecurityPolicy and WS-MetadataExchange are used. The identity management applied the described decoupled claim-based SSO architecture. In our experiment, we assume that a user authenticates himself by supplying his username and password token to the identity provider (IdP) (see Figure 2), who will create a SAML-Assertion describing the authentication status of the user and hand out a symmetric key which will be used to secure the connection to the Relying Party (RP). Finally, this assertion token is used to call a policy-constrained Web service.

Furthermore, we implemented ourselves a SSO system in the Ubipol project (see section V) using the WS-* technologies. In this project, we used the same architecture as already described.

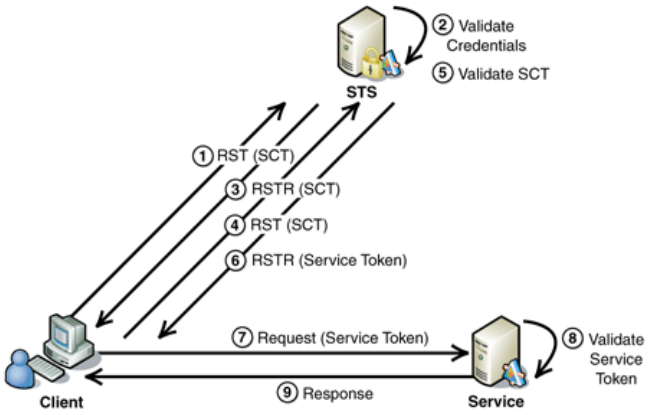


Figure 2: WS secure conversation

IV. Experimental Results

We have analyzed the traffic which was generated by decoupled claim-based scenarios. First of all, a simple application has been employed, which makes use of the *Metro* Web service stack implementation (*Metro 2.1.1*). *WSIT* is part of *Metro* and provides methods for creating reliable, transactional and secure Web services using standards specified by the standardization group OASIS.

Second, we created our own implementation and compared the traffic of both scenarios.

No compression is used for the traffic analysis in order to inspect the distribution of the traffic regarding the Web service security standards. We will distinguish between the first and second (or further) Web service call made by the client. All packets which are necessary for succeeding the request are regarded as one package. This may involve additional traffic to other recipients, in particular IdP. All Web service security relevant XML elements were considered and their generated traffic has been assigned to the related standard based on the qualified name which itself and only its content from the same standard are part of the counted size.

In the first scenario, the first user's Web service call is accompanied by 4.5 times more traffic than the further ones (see Figure 3). This is caused by bootstrapping the Web service, which contains the policy and *WSDL* retrieval, but also the token request to the IdP needs to be done. The former aspect explains, why *WS-Trust* and *WS-MetadataExchange* are not represented in following Web service calls, because they are part of the initial token exchange process.

All further requests to RP require less traffic, because all relevant keys have been established and put into a security context. As defined in *WS-SecureConversation*, derived keys are used for protecting the transported content. In the first request, the security standards use 85% of the traffic. This reduces in any further requests to 33%.

In the second scenario, where we analyzed the traffic of our own implementation, the traffic of the first request is also much higher than in any subsequent requests. Anyhow, the overall package size is much smaller since our implementation does not use any *Metadata-Exchange* and also the transferred data of *WS-Security* is slightly lower. Therefore the ratio between security data and payload transported is much better. In the first request, the security data creates 54% of the total package size and in further requests lowers to only 8% (see Figure 4).

Even though modern mobile networks (3G and 4G) solve bandwidth problems, high-latency issues still exist. In the bootstrapping process the client request is forwarded to the IdP, which implies an additional expensive Web service call. When dealing with low traffic Web services the initial waiting time may be inconvenient due to the collateral traffic and the mentioned further forward to the IdP.

V. Security in Mobile Based ePArticipation

eParticipation refers to the ICT mechanisms for the citizens to express their opinions in order to influence political, economic or social decisions. Recently, the rapid advance in mobile computing technologies also facilitated the emergence of mParticipation (mobile participation) to allow citizens to be involved in Policy Making Processes (PMPs) even on the move. The UbiPOL project [10] aims to

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

develop a new governance model in which citizens can participate in policy making processes in the middle of their everyday life overcoming spatial and time barriers. The core of the governance model is a ubiquitous participation platform that motivates its users to be involved in PMPs. Depending on the status of the relevant policies, the citizen may add his own opinion on his handheld device which will deliver the opinion to the opinion base of the relevant government agency. The collected citizen opinions for each site objects will be connected with relevant policy objects to be used for policy making process.

One of the main objectives of UbiPOL is to develop a framework that

1. Ensures citizens privacy in filtering citizen opinions
2. Secures the communication between the mobile device and the Ubipol platform
3. Ensures the anonymity of the user in case of opinion casting. This means the association between the opinions records and the user identity must remain unknown
4. Prevents multiple opinions casting
5. Manages the Ubipol platform users identities and regulates the access to it according to the user role

In the UbiPOL project, Fraunhofer Fokus is in charge of the items (2), (3), (4) and (5). Contrary to the work discussed in the literature about anonymous voting ([11], [12]) where some complex algorithms are suggested, we came out in this project with a simple Web-based architecture that addresses the above issues while taking into account the limitations of the mobile devices. The architecture is already specified and is being implemented and tested on Android capable mobile devices. For the time being, simple credentials (user name, password) are used for authentication and voting tickets acquisition. In the near future, we will enhance the framework with the use of the new German eID cards.

VI. CONCLUSION AND FUTURE WORK

This paper provided an introduction in state of the art specifications for securing Web services. Then a scenario has been described where these technologies were used in order to implement a single sign-on system. Based on two implementations (one using NetBeans and another one developed by the authors), a traffic analysis has been performed and showed that WS-* technologies can produce high overhead (more than 85% of the entire bandwidth) in low-traffic Web service especially when using *Metadata-Exchange*. In the future, we will further optimize our realization for mobile devices by investigating the security standards and group them into profiles that can be used for appropriate scenarios.

ACKNOWLEDGMENT

This work was achieved within the UbiPol project [10]. UbiPol is supported by the European Community under the FP7 ICT Work Programme (call: ICT-2009.7.3 (a)).

REFERENCES

[1] SAML, "Security Assertion Mark-up Language". Link http://saml.xml.org/saml-specifications, Access: March 2012

[2] OASIS: "Advancing open standards for the information society", link: http://www.oasis-open.org/, Access: March 2012

[3] WS-Trust, link: http://docs.oasis-open.org/ws-sx/ws-trust/v1.4/ws-trust.html, Access: March 2012

[4] KERBEROS: The Network Authentication Protocol, link: http://Web.mit.edu/kerberos/, Access: March 2012

[5] OpenID, link: http://openid.net/foundation/, Access: March 2012

[6] SOAP messages, link: http://www.w3.org/TR/soap/, Access: March 2012

[7] W3S, link: http://www.w3.org/, Access: March 2012

[8] Apache Axis 2, link: http://axis.apache.org/axis2/java/core/, Access: March 2012

[9] Metro Glassfish, link: http://en.wikipedia.org/wiki/GlassFish_Metro, Access: March 2012

[10] The Ubipol project, link: http://www.ubipol.eu/, Access: March 2012

[11] A. Y. Lindell, "Anonymous Authentication", Aladdin Knowledge Systems Inc, Bar-Ilan University, Israel, 2006. Link: http://www3.safenet-inc.com/blog/pdf/AnonymousAuthentication.pdf, Access: March 2012

[12] K. Sako; S. Yonezawa; and I. Teranishi; "Anonymous Authentication: For Privacy and security", NEC Journal of Advanced Technology, Special Issue on Security for Network Society, Vol. 2, No. 1, 2005.

[13] Project U-Prove, link: http://www.fokus.fraunhofer.de/de/fokus_testbeds/secure_eidentity-lab/projekte/u_prove/index.html, Access: March 2012

[14] Is Microsoft's U-Prove the answer to better online privacy, link: http://www.networkworld.com/community/blog/microsofts-u-prove-answer-better-online-priva, Access: March 2012

[15] NetBeans IDE integration, link: http://glassfish.java.net/public/netbeans/index.html

[16] NetBeans, link: http://en.wikipedia.org/wiki/NetBeans

[17] Apache CXF, link: http://cxf.apache.org/, Access: March 2012

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Feature	Axis 1.x	Axis2	CXF	Glue	JBossWS	XFire	Metro@GlassFish	OracleAS 10g
WS-Addressing	X	X	X	X	X	X	X	with BPEL
WS-Atomic Transaction	X	X					X	
WS-Business Activity		X						
WS-Coordination	X	X					X	
WS-Eventing		X			X			
WS-Metadata Exchange		X [10]	X				X	
WS-Notification	X	X [12]	X	?		?		
WS-ReliableMessaging	X	X	X				X	
WS-Policy		X	X				X	X
WS-Secure Conversation		X	X				X	
WS-Security Policy		X	X				X	
WS-Security	X	X	X	X	X	X	X	X
WS-Trust		X	X				X	
WS-Transfer		X						
WSDL 1.1 Support	X	X	X	X	X	X	X	X
WSDL 2.0 Support		X						

Table 1: Comparison of the different framework

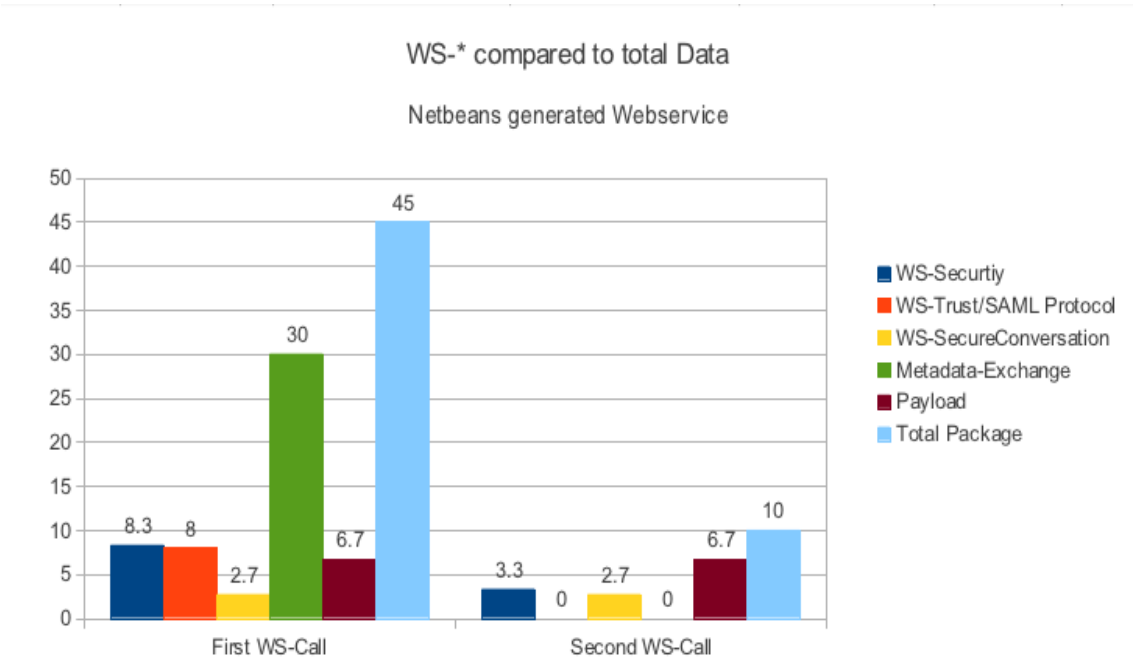


Figure 3: WS-* Compared total data

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

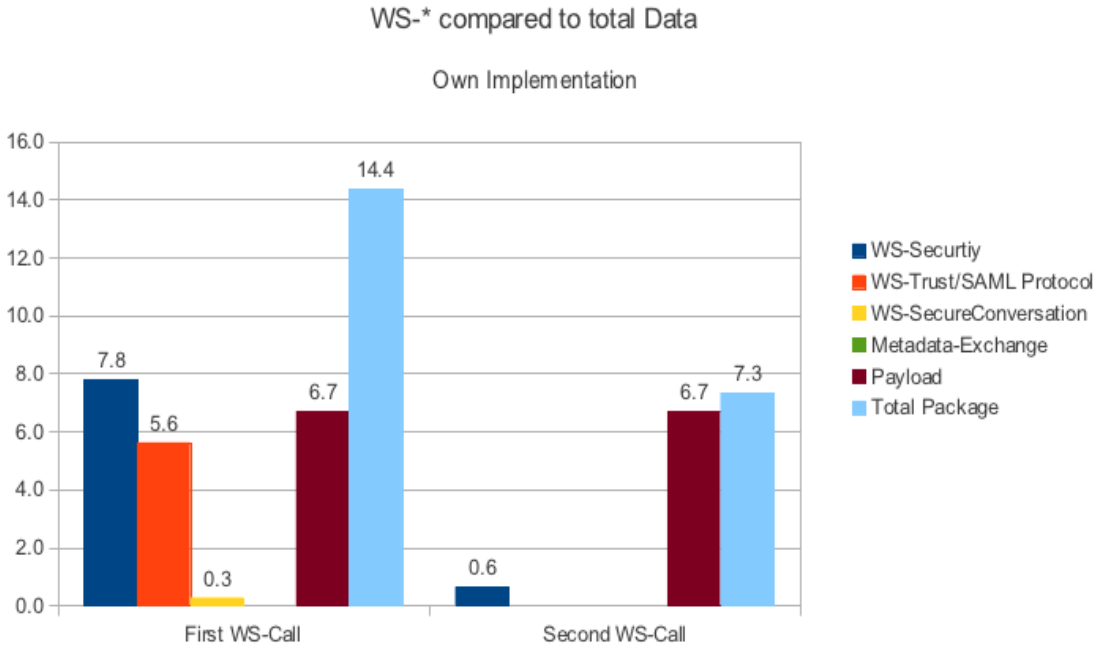


Figure 4: Our WS-* implementation compared to total data

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

EXIT Charts Analysis for Turbo-TCM Schemes Using Non-Binary RSC Encoders

Calin Vladeanu, Alexandru Martian
Telecommunications Department
University Politehnica of Bucharest
Bucharest, Romania
Email: calin@comm.pub.ro

Safwan El Assad
IETR Laboratory
UMR CNRS 6164*, Image team - site of Nantes*
École d'Ingénieurs de l'Université de Nantes
Nantes, France
Email: safwan.elassad@univ-nantes.fr

***Abstract*—Recently, a new family of recursive systematic convolutional (RSC) encoders over Galois field GF$\left(2^N\right)$ was introduced. The present paper considers the parallel turbo trellis coded modulation (TTCM) scheme using these non-binary encoders as constituent codes. Besides operating over a higher order Galois field, these encoders use a non-linear function, the left-circulate (LCIRC) function, to control the encoder states. It is shown that these rate-$(N-1)/N$ GF$\left(2^N\right)$ RSC-LCIRC encoders offer the same performances in terms of minimum Euclidian distance, effective length, and product distance, as compared to corresponding binary encoders. Moreover, these RSC-LCIRC encoders are requiring less memory usage. Extrinsic information transfer (EXIT) charts are used to analyze the convergence of the proposed punctured TTCM schemes, with 8-PSK and 16-QAM modulations, symbol-by-symbol log-MAP decoding algorithm, when transmitting over AWGN and Rayleigh non-selective fading channels. EXIT charts show some improvements in terms of TTCM decoding schemes convergence for the RSC-LCIRC codes as compared to their binary counterparts.**

***Keywords*-EXIT charts; GF$\left(2^N\right)$ encoders; Punctured Turbo TCM; Symbol-by-Symbol log-MAP; LCIRC**

I. INTRODUCTION

In the latter years, nonlinear functions proved to be suitable to implement several telecommunications systems blocks, such as pseudo-random number generation, spreading for code-division multiple access systems, and encryption. Among these nonlinear functions the ones operating in finite precision are more suitable for digital implementations. In [1], Frey proposed a nonlinear digital infinite impulse response (IIR) filter for secure communications. The Frey filter contains a nonlinear function named left-circulate function (LCIRC), which provides the chaotic properties of the filter. The above mentioned work considered the Frey encoder as a digital filter, operating over Galois field GF$\left(2^N\right)$. In [2] it was demonstrated that the Frey encoder with finite precision (wordlength of N bits) presented in [1] is a recursive convolutional encoder operating over GF$\left(2^N\right)$. New methods for enhancing the performances of the phase shift keying - trellis-coded modulation (PSK-TCM) transmissions over a noisy channel, using the recursive convolutional LCIRC (RC-LCIRC) encoders, were proposed in [3]. These encoders follow the rules proposed by Ungerboeck [4] for defining optimum TCM by proper set partitioning for channels with additive white Gaussian noise (AWGN). When assuming a Rayleigh non-selective fading channel, other designing criteria can be used to optimize the TCM system performances, such as minimum effective length and product distance [5]. Turbo coded schemes were developed as well for the TCM schemes [6], [7]. A family of nonlinear encoders for turbo TCM (TTCM) scheme was proposed in [8]. The present paper aims to extend the performances analysis for these nonlinear TTCM schemes. The extrinsic information transfer (EXIT) chart is an important tool for visualizing the exchange of the extrinsic information between constituent decoders in a turbo receiver scheme [9]. The EXIT chart was also applied to turbo TCM (TTCM) schemes to depict the decoding trajectory, allowing the prediction of bit error rate (BER) waterfall and BER floor regions [10], [11]. Therefore, the EXIT chart can be used as a tool in the design of TTCM schemes [12].

In the present work, the analysis of the TTCM scheme from [8] is extended by considering also the quadrature amplitude modulation (QAM) and a Rayleigh fading channel. Moreover, the EXIT charts are presented to underline the convergence behavior of these schemes.

The paper is organized as follows. Section II is presenting the recursive systematic convolutional LCIRC (RSC-LCIRC) encoder operating over Galois field GF$\left(2^N\right)$ and the optimum set partitioning for two-dimentional TCM schemes. The minimum Euclidian distance, minimum effective length and product distance are estimated for the AWGN channel, and the Rayleigh fading channel, respectively. In Section III, a parallel TTCM transmission scheme using RSC-LCIRC component encoders with symbol puncturing is presented. A symbol-by-symbol log-MAP algorithm is used for the iterative detection. EXIT charts are plotted in Section IV to compare the convergence of different TTCM decoding schemes for the punctured 8-PSK and 16-QAM TTCM transmissions. Finally, the conclusions are drawn in Section V.

II. OPTIMUM RSC-LCIRC ENCODER FOR TCM SCHEMES

In this section, a new family of RSC encoders operating over Galois field GF$\left(2^N\right)$ and their use for optimum TCM

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

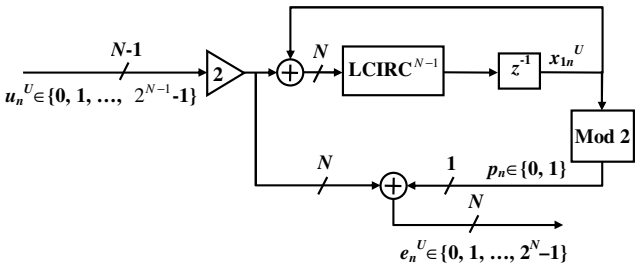


Figure 1. Rate $(N-1)/N$ optimum GF(2^N) RSC-LCIRC encoder.

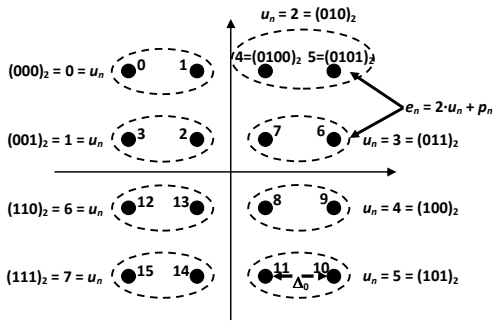


Figure 2. Set partitioning for punctured 16-QAM TTCM.

schemes are presented. The main component of the RSC encoder presented in the sequel is the nonlinear LCIRC function introduced by Frey in [1] for chaotic encryption. The use of the LCIRC function for channel encoding was considered for the first time in [2]. Optimum encoders using the LCIRC function for TCM schemes were introduced in [3]. However, despite being characterized by optimum Euclidian distances, these recursive convolutional encoders are non-systematic. Therefore, the coding features of these non-systematic encoders are not fully exploited in turbo schemes. In this section, we introduce a new encoder operating over Galois field GF$\left(2^N\right)$, using the LCIRC function, which is systematic, i.e., the encoder output value specifies explicitly the input value. Let us denote by N the wordlength used for binary representation of each sample. The LCIRC function performs a bit rotation by placing the most significant bit to the least significant bit, and shifting the other $N-1$ bits one position to a higher significance.

The block scheme for a rate $(N-1)/N$ RSC-LCIRC encoder, using one delay element and the LCIRC function is presented in Fig. 1. For each moment n, u_n represents the input data sample, x_{1n} denotes the delay output or the encoder current state, and e_n is the output sample. The superscript U denotes that all the samples are represented in unsigned N bits wordlength, i.e., $u_n^U \in \left[0, 2^{N-1}-1\right]$, $e_n^U \in \left[0, 2^N-1\right]$. The rate for the encoder in Fig. 1 is the ratio between the input wordlength $N_{in} = N-1$ and the output wordlength N, i.e., $R = N_{in}/N$ [2]. LCIRC^{N-1} represents the LCIRC function application for $N-1$ times consecutively. Both adders and the multiplier are modulo-2^N operators. The modulo-2 block extracts the least significant bit, denoted by p_n, from the encoder current state value, x_{1n}. Therefore, p_n is the parity bit for the systematic rate $(N-1)/N$ encoder. It is important to demonstrate that the encoder is systematic, i.e., the encoder output binary representation codeword e_n^U includes the representation codeword of the encoder input u_n^U. Hence, the output is obtained by shifting the $N-1$ bits of the input representation codeword by one position to a higher significance, and adding the parity bit p_n to the least significant position, a position that was left empty, inside the N bits output codeword, by the previously mentioned shifting. The one position shifting presented above is equivalent to a multiplication by 2 in the GF$\left(2^N\right)$ field. Therefore, the encoder output value e_n^U is given by the following GF$\left(2^N\right)$ equation:

$$e_n^U = 2 \cdot u_n^U + p_n = 2 \cdot u_n^U + x_{1n}^U \bmod 2 \qquad (1)$$

The trellis complexity of the codes generated with the scheme in Fig. 1 increases with the wordlength N, because the number of trellis states grows exponentially with the output wordlength, i.e., 2^N, while the number of transitions originating from and ending in the same state grows exponentially with the input wordlength, i.e., 2^{N-1}.

The set partitioning for the punctured TTCM scheme, which optimizes the initialization of the a priori information for the first decoder, during the first iteration, was introduced in [6], and has two features. First of all, the set partitioning follows the Ungerboeck optimum set partitioning rules from [4], and secondly, the constellation points associated to the same group of $N-1$ systematic information bits, i.e., to the same input symbol u_n^U, but differing in the least significant bit, i.e., the parity bit p_n, should be placed at the minimum distance in the set, $\Delta_{0,2^N-\text{ary modulation}}$. Following these two requirements, the optimum set partitioning rule for 16-QAM is depicted in Fig. 2. The first feature maximizes the minimum Euclidian distance of the component TCM code, while the second feature minimizes the distance between elements of the subsets associated to identical systematic bits, denoted by ovals in Fig. 2, for the global punctured TTCM code.

It can be easily demonstrated that the minimum Euclidian distance for the 2^N-ary TCM component encoder presented in Fig. 1, using the properly partitioned constellations, has the following expression:

$$d^2_{E,R=\frac{N-1}{N}} = \begin{cases} 2\Delta^2_{1,2^N-\text{PSK}} + \Delta^2_{0,2^N-\text{PSK}}, \text{for PSK} \\ 5\Delta^2_{0,2^N-\text{QAM}}, \text{for QAM} \end{cases} \qquad (2)$$

In Table I, some values of the minimum distance of the TCM encoder in Fig. 1 are presented for different values of N, and for the PSK and QAM constellations, respectively. The associated coding rates are presented in the second column. It can be easily noticed from (2) that all the rate

Table I
MINIMUM 2^N-ARY TCM DISTANCES AS FUNCTION OF N FOR OPTIMUM GF(2^N) RSC-LCIRC ENCODERS.

N	*R*	**Modulation**	d_E^2	l'_m	d_p^2
2	1/2	QPSK, 4-QAM	10	3	32
3	2/3	8-PSK	≈ 4.5858	2	8
4	3/4	16-PSK	≈ 1.3238	2	≈ 1.1716
4	3/4	16-QAM	2	2	1.28

$(N-1)/N$, for any N value, the RSC-LCIRC encoders have the same minimum distances as the corresponding binary optimum encoders [6], [7]. When transmitting over a Rayleigh non-selective fading channel the TCM system performances depend mainly on two parameters, i.e., the minimum effective length, and the minimum product distance [5]. We call the minimum effective length l_m the length of the shortest path pair of encoder output values (x_n, x'_n). Among these paths of length l_m there is one having the smallest product distance $d_p^2 = \Pi_{n=1, x_n \neq x'_n}^{l_m} \mid x_n - x'_n \mid^2$. The values of these parameters are presented in the last two columns from Table I, for the same RSC-LCIRC codes. Again, the RSC-LCIRC encoders have the same values for minimum effective length and product distance, as their binary counterparts [5]. However, the GF$\left(2^N\right)$ RSC-LCIRC encoders are less complex than the corresponding binary encoders in terms of memory usage. The memory size of the binary encoders increases logarithmically with the number of states in the trellis, while the GF$\left(2^N\right)$ RSC-LCIRC encoders include only one delay element, no matter what the trellis complexity is. As another advantage of these encoders, we can also mention the Euclidian distance compact expression (2) as a function of N.

III. RSC-LCIRC ENCODER IN TURBO-TCM SCHEME

Fig. 3 shows the turbo TCM transmitter for 2^N-ary modulation. The information 2^{N-1}-ary symbol sequence u_n and its block-wise interleaved version $u_n^{(i)}$ are fed into two identical component encoders RSC-LCIRC$_1$ and RSC-LCIRC$_2$ of rate $(N-1)/N$. The encoders' outputs are selected alternatively and mapped into 2^N-ary modulated symbol sequence x_n. The output of the bottom encoder is deinterleaved according to the inverse operation of the interleaver. This ensures that at the input of the symbol selector, the $N-1$ information bits from the 2^{N-1}-ary input symbol, partly defining the encoded 2^N-ary symbols of both the upper and lower input, are identical [6], [7]. Therefore, if the selector is switched on a symbol base, the mapper output is a punctured version of the two encoded sequences, and the $N-1$ information bits appear only once, mapped in a single transmitted symbol selected either from e_{1n} sequence or from e_{2n} sequence. Nevertheless, the remaining parity

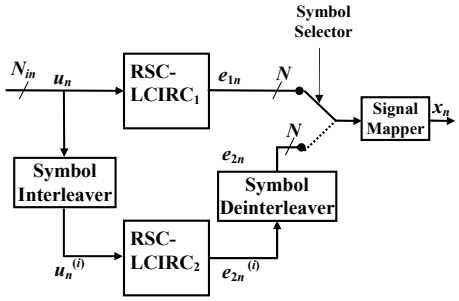


Figure 3. TTCM transmitter with RSC-LCIRC encoders and symbol puncturing.

bit carried by the transmitted symbol is taken alternatively from the upper and lower encoder. Hence, the overall coding rate for the scheme in Fig. 3 is $(N-1)/N$. The 2^N-levels modulated symbol sequence is transmitted over a noisy and non-selective fading channel. The received signal over the n-th symbol interval is given by:

$$y_n = h_n x_n + w_n \tag{3}$$

where w_n is an AWGN sequence with $E\left[|w_n|^2\right] = N_0$, and x_n denotes the 2^N-levels symbol value mapped from the encoders output sequences (e_{1n}, e_{2n}) by puncturing over the n-th symbol interval. The coefficient h_n is the path gain from transmit antenna to receive antenna, having a Rayleigh distribution. The path gains are modeled as the absolute part of samples of independent complex Gaussian random variables with variance 0.5 per real dimension. The wireless channel is assumed to be quasistatic, i.e., the path gain value is constant over a group of symbol intervals, and varies from one group of symbols to another. The receiver structure including two component decoders that use the symbol-by-symbol log-MAP algorithm was introduced in [6]. The decoding process is similar to the binary turbo decoding, except that the symbol probability is used as the extrinsic information rather than the bit probability [6], [7]. The log-MAP decoder computes the log likelihood ratio (LLR) for each group of information bits transmitted at the n-th symbol interval u_n, embedded in the 2^{N-1}-ary input symbol taking one of the integer values $j \in \{0, 1, ..., 2^{N-1}-1\}$ as [7]

$$L\left(u_n = j\right) = \ln \frac{P\left(u_n = j|\mathbf{y}\right)}{P\left(u_n = 0|\mathbf{y}\right)} \tag{4}$$

where **y** is the received signal vector. The symbol-by-symbol log-MAP decoder operates on an L symbols block basis. Hence, in all equations the symbol time variable n takes values between 1 and L. We assume that the receiver has perfect side information on the path gains (h_n). The input symbol j with the largest LLR in (4) is chosen as the hard decision output.

A very important tool for the iterative decoding performances analysis consists in the EXIT chart, which de-

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

scribes the extrinsic mutual information exchange between constituent decoders. A complexity efficient method for generating the symbol-based EXIT charts from symbol-based a posteriori probabilities (APPs) was proposed in [11]. The expression for the average extrinsic information $I_E(u)$, estimated at the output of the decoder for the input symbol vector u, is given by [11]:

$$I_{E,D}(u) = N - 1 + \frac{1}{L}\sum_{n=1}^{L} \mathbf{E}\left[\sum_{i=1}^{2^{N-1}} e_D\left(u_n^{(i)}\right) \cdot \log_2\left(e_D\left(u_n^{(i)}\right)\right)\right] \quad (5)$$

where L is the number of information symbols in the decoded block, $N-1$ is the number of information bits per input symbol, $u_n^{(i)}$ is the presumed transmitted information symbol at time instant n for $i \in \{1, 2, ..., 2^{N-1}\}$, and $e(\cdot)$ is the extrinsic probability. The expectation $\mathbf{E}[\cdot]$ can be approximated by simple time-averaging of the extrinsic probabilities of the information symbol. We propose to approximate the extrinsic probability as the normalized joint extrinsic and systematic information of the log-MAP decoder:

$$e_D\left(u_n^{(i)}\right) \approx \frac{\exp\left(L_{D,es}(u_n = i)\right)}{\sum_{i=1}^{2^{N-1}} \exp\left(L_{D,es}(u_n = i)\right)} \quad (6)$$

In equations (5) and (6), D denotes the decoder number, i.e., $D \in \{1, 2\}$. On the other hand, computing an exact value of the extrinsic probability requires that the systematic and parity parts of the channel observation variables are independent. Thus, equation (6) represents only an approximation of the true extrinsic information. The average a priori information $I_A(u)$ is computed in a similar manner. The EXIT chart is obtained by representing, on the same diagram, the decoder 1 transfer characteristic, i.e., $I_{E,1} = T(I_{A,1})$, and the decoder 2 transfer characteristic, i.e., $I_{E,2} = T(I_{A,2})$, for a constant E_b/N_0 value. The axes of the decoder 2 transfer characteristic are swapped.

IV. Simulations Results

The TTCM scheme presented in Section III using the RSC-LCIRC encoders presented in Section II was tested for 8-PSK and 16-QAM by means of simulations over an AWGN non-faded channel, and over an AWGN and non-selective fading channel, respectively. Both component encoders in the TTCM scheme are identical rate-$2/3$ RSC-LCIRC encoders for 8-PSK, and rate-$3/4$ for 16-QAM, respectively. The modulation is using the optimum set partitioning for the punctured TTCM scheme as presented in Section II. The symbol interleavers used for simulations are pseudo-random and operate independently on even and odd positions [6]. The symbol-by-symbol log-MAP decoding algorithm is used in the receiver. The decoding convergence is investigated for several E_b/N_0 values, where E_b is the signal energy per bit and N_0 is one-sided power spectral density of the AWGN noise. The interleaver block includes 1024 symbols. As references, we considered the

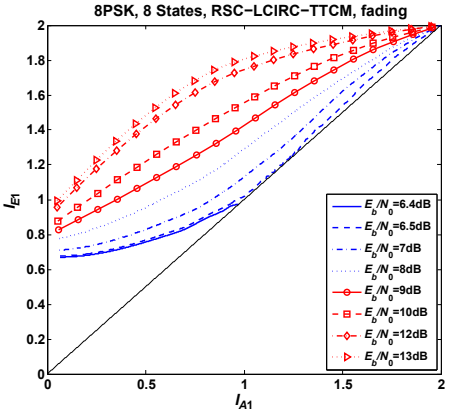
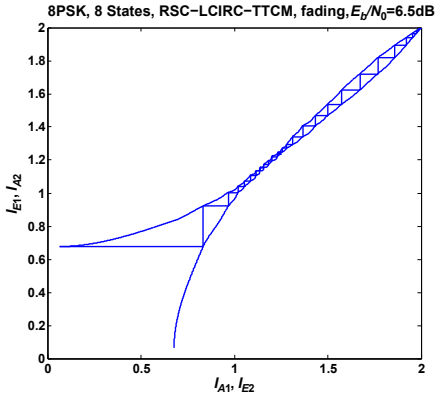


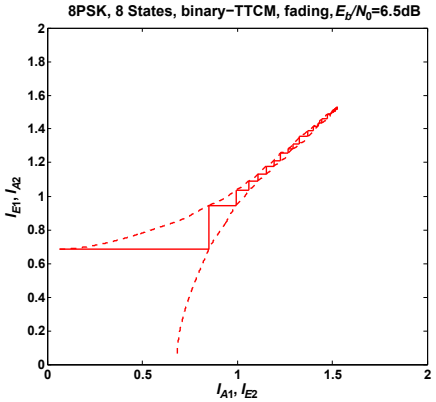
Figure 4. Extrinsic information transfer characteristic of 8PSK-RSC-LCIRC-TTCM decoder for a fading channel.

corresponding optimum binary encoders. Therefore, for rate-$2/3$ 8-PSK TTCM scheme the optimum binary encoder with 8 states is given by the the generator polynomials, represented in octal notation $[11, 02, 04]$ [6]. The rate-$3/4$ optimum binary encoder with 16 states, considered as reference, was determined in [7] for 16-QAM TTCM with the generator polynomials $[23, 35, 33, 37]$. In the following, EXIT charts are contrived for 8PSK and 16QAM schemes presented above, using a simulation procedure described in [11]. These EXIT charts are relevant for the TTCM decoder convergence analysis, revealing important features, such as BER turbo cliff and BER floor regions. The average extrinsic information $I_E(u)$ and the average a priori information $I_A(u)$ are estimated using equation (5), assuming that the extrinsic probability is approximated with (6). Fig. 4 shows the extrinsic information transfer characteristic of 8PSK-RSC-LCIRC-TTCM decoder for a fading channel, assuming a variable E_b/N_0. The considered scenario assumes the transmission over a non-selective Rayleigh fading channel, LCIRC encoded 8PSK TTCM scheme, and a blocklength of 2^{14} symbols. Analyzing the curves in Fig. 4 one can easily notice that above $E_b/N_0 = 6.4$ dB, the average decoding trajectory obtained from real simulations shows convergence. In Fig. 5(a), the EXIT chart for $E_b/N_0 = 6.5$ dB is depicted. This EXIT chart plots the *bottleneck region* with the decoding trajectory just managing to pass through a narrow tunnel, which corresponds to the BER waterfall region. In Fig. 5(a), the convergence is almost reached after 30 iterations. The EXIT chart obtained under the same assumptions, for the corresponding binary encoder, is plotted in Fig. 5(b). It is clear that for the binary case decoder the trajectory gets stuck as compared to the LCIRC decoder, for $E_b/N_0 = 6.5$ dB, assuming the same number of iterations. In Fig. 5(c) the EXIT chart corresponding to the *wide-open region* is depicted. The scenario is identical to the previous one, but for $E_b/N_0 = 13$ dB. This region is related to

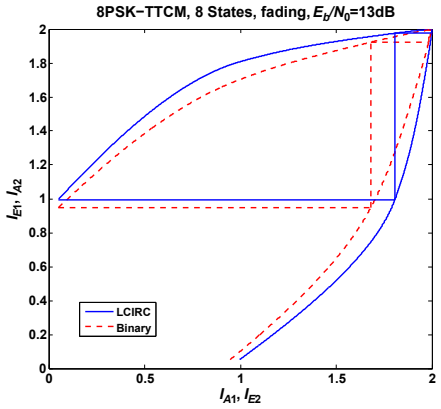
Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1



(a)



(b)



(c)

Figure 5. EXIT charts for 8PSK-TTCM over a fading channel. (a) LCIRC, $E_b/N_0 = 6.5$ dB; (b) Binary, $E_b/N_0 = 6.5$ dB; (c) $E_b/N_0 = 13$ dB.

the BER floor region. The trajectories for both binary and LCIRC decoders are depicted in Fig. 5(c). Again, the LCIRC decoder outperforms its binary counterpart due to the wider opening of the EXIT chart. In this case, the convergence is almost reached after 3 iterations.

Fig. 6 shows the extrinsic information transfer characteristic of 16QAM-RSC-LCIRC-TTCM decoder for a non-fading channel, assuming a variable E_b/N_0. A blocklength of 2^{14} symbols was considered. Analyzing the curves in Fig. 6, one can easily notice that above $E_b/N_0 = 5.4$ dB, the average decoding trajectory shows convergence. In Fig. 7(a), the EXIT chart for $E_b/N_0 = 5.4$ dB is depicted. This EXIT chart plots the *bottleneck region*, which corresponds to the BER waterfall region. In Fig. 7(a), the convergence is almost reached after 10 iterations. The EXIT chart obtained under the same assumptions, for the corresponding binary encoder, is plotted in Fig. 7(b). It is clear that for the binary case decoder, the trajectory is better than the LCIRC decoder one. In fact, the binary decoder needs less iterations (only 7) to converge, and the EXIT diagram is wider. In Fig. 7(c) the EXIT chart for $E_b/N_0 = 6$ dB, corresponding to the *wide-open region*, is depicted. In this case, the LCIRC decoder outperforms its binary counterpart, due to the wider opening of the EXIT chart. Moreover, the convergence is almost reached after 4 iterations. As a conclusion, the LCIRC outperforms the binary case for $E_b/N_0 > 5.8$ dB, while the opposite happens for lower values of E_b/N_0.

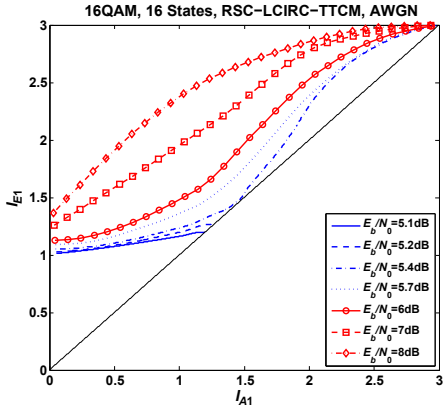
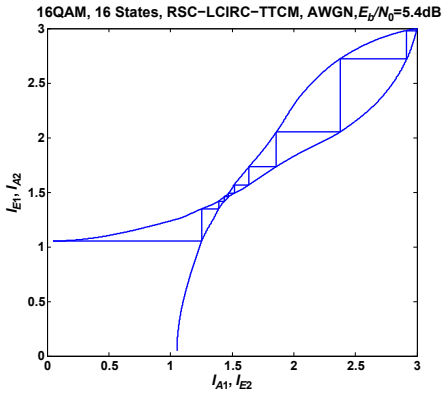


Figure 6. Extrinsic information transfer characteristic of 16QAM-RSC-LCIRC-TTCM decoder for a non-fading channel.

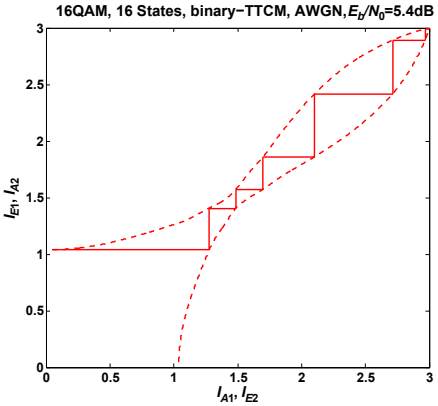
V. Conclusion

It was demonstrated that using the LCIRC function, efficient RSC encoders over GF$\left(2^N\right)$ can be designed for punctured TTCM transmissions. A generalized 1-delay GF(2^N) RSC encoder scheme using LCIRC was defined, for any possible encoding rate of $(N-1)/N$. It was shown that LCIRC-based encoders offer at least the same performances as conventional binary encoders for the rate-$(N-1)/N$ schemes. EXIT charts were provided to demonstrate their convergence properties. In perspective, we intend to extend the GF(2^N) RSC encoders design using different nonlinear functions.

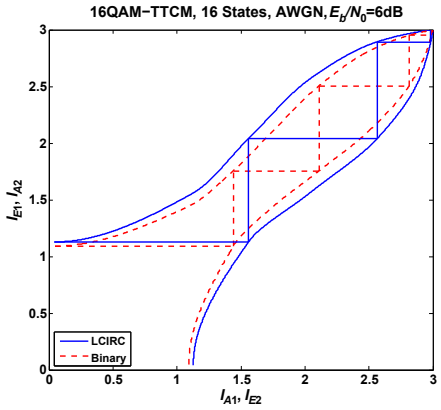
Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1



(a)



(b)



(c)

Figure 7. EXIT charts for 16QAM-TTCM over a non-fading channel. (a) LCIRC, $E_b/N_0 = 5.4$ dB; (b) Binary, $E_b/N_0 = 5.4$ dB; (c) $E_b/N_0 = 6$ dB.

ACKNOWLEDGMENT

This work was supported in part by the Romanian contract POSDRU/89/1.5/S/62557 and by the Romanian UEFISCSU PN-2 RU-TE Project no. 18/12.08.2010.

REFERENCES

[1] D. R. Frey, "Chaotic digital encoding: An approach to secure communication," *IEEE Trans. Circuits and Systems - II: Analog and Digital Signal Processing*, vol. 40, pp. 660-666, Oct. 1993.

[2] C. Vladeanu, S. El Assad, J.-C. Carlach, and R. Quéré, "Improved Frey Chaotic Digital Encoder for Trellis-Coded Modulation," *IEEE Trans. Circuits and Systems - II*, vol. 56, pp. 509-513, Jun. 2009.

[3] C. Vladeanu, S. El Assad, J.-C. Carlach, R. Quéré, I. Marghescu, "Recursive GF(2^N) Encoders Using Left-Circulate Function for Optimum PSK-TCM Schemes," *Signal Processing*, vol. 90, pp. 2708-2713, Sep. 2010.

[4] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Information Theory*, vol. IT-28, pp. 55-67, Jan. 1982.

[5] C. Schlegel and D.J. Costello, "Bandwidth Efficient Coding for Fading Channels: Code Construction and Performance Analysis," *IEEE Trans. on Sel. Areas in Comm.*, vol. 7, pp. 1356-1368, Dec. 1989.

[6] P. Robertson and T. Worz, "Bandwidth-Efficient Turbo Trellis-Coded Modulation using Punctured Component Codes," *IEEE Trans. on Sel. Areas in Comm.*, vol. 16, pp. 206-218, Feb. 2000.

[7] B. Vucetic and J. Yuan, *"Turbo Codes: Principles and Applications*," Springer, 2000.

[8] C. Vladeanu and S. El Assad, "Punctured 8-PSK Turbo-TCM Transmissions Using Recursive Systematic Convolutional GF(2^N) Encoders," in *Proc. 19th European Signal Conf. - EUSIPCO 2011*, Barcelona, Spain, Aug. 29 - Sept. 2, 2011, pp. 111-115.

[9] S. ten Brink, "Convergence Behavior of Iteratively Decoded Parallel Concatenated Codes," *IEEE Trans. on Comm.*, vol. 49, pp. 1727-1737, Oct. 2001.

[10] H. Chen and A. Haimovich, "EXIT Charts for Turbo Trellis-Coded Modulation," *IEEE IEEE Commun. Lett.*, vol. 8, pp. 668-670, Nov. 2004.

[11] J. Kliewer, S. X. Ng, and L. Hanzo, "Efficient Computation of EXIT Functions for Nonbinary Iterative Decoding," *IEEE Trans. on Comm.*, vol. 54, pp. 2133-2136, Dec. 2006.

[12] Soon Xin Ng, O.R. Alamri, Li Yonghui, J. Kliewer, L. Hanzo, "Near-Capacity Turbo Trellis Coded Modulation Design Based on EXIT Charts and Union Bounds," *IEEE Trans. on Comm.*, vol. 56, pp. 2030-2039, Dec. 2008.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

PAPR Reduction of OFDM Signals using Active Constellation Extension and Tone Reservation Hybrid Scheme

Eugen-Victor Cuteanu
Communication Department
Politehnica University, Faculty of Electronics and Telecommunications
Timisoara, Romania
victor.cuteanu@gmail.com

Alexandru Isar
Communication Department
Politehnica University, Faculty of Electronics and Telecommunications
Timisoara, Romania
alexandru.isar@etc.upt.ro

***Abstract*—The Orthogonal Frequency Division Multiplexing is one of the widely used modulation techniques in the present broadband wireless technology. The opportunities and challenges of this modulation technique are derived from its native advantages and disadvantages. One of the main problems is the high peak-to-average power ratio of transmission signal due to the superposition of many subcarriers. This paper presents a new hybrid peak-to-average power ratio reduction technique, which combines an active constellation extension method with a tone reservation method. The paper presents the performance and advantages of the mixed technique and compares it with other existing methods. The simulations shown that the proposed technique realizes an increased peak-to-average power ratio reduction compared to component methods with similar parameters.**

Keywords-OFDM; PAPR; Active Constellation Extension; Tone Reservation

I. INTRODUCTION

The Orthogonal Frequency Division Multiplexing (OFDM) is one of the most efficient and popular modulation techniques used in broadband wireless communication systems like Worldwide Interoperability for Microwave Access (WiMAX), Terrestrial Digital Video Broadcast (DVB-T), or wireline systems like Asymmetric Digital Subscriber Line (ADSL). One of the main practical issues of the OFDM is the Peak-to-Average Power Ratio (PAPR) of the transmitted signal. This high PAPR occurs because of the time-domain superposition of the many data subcarriers which compose the OFDM signal. Due to the large number of subcarriers, the resulting time-domain signal exhibits Rayleigh-like characteristics and large time-domain amplitude variations. These large signal peaks require the high power amplifiers (HPA) to support wide linear dynamic range.

Higher signal level causes non-linear distortions leading to an inefficient operation of HPA causing intermodulation products resulting unwanted out-of-band power. In order to reduce the PAPR of OFDM signals, many solutions have been proposed and analyzed. The efficiency of these methods can be evaluated considering their characteristics of non-linearity, amount of processing and size of side information needed to be sent to receiver.

The class of linear methods is represented by approaches like partial transmit sequence (PTS) [2], selective mapping (SLM) [1], and tone reservation (TR) [5].

In the SLM method, based on a set of predefined phase arrays, several vector rotations of the original frequency domain OFDM signal are performed. For each variant obtained by rotations, the corresponding PAPR is evaluated. The variant with the lowest PAPR is chosen for the transmission.

A similar approach is applied in case of PTS method, where the N complex values representing the OFDM signal symbols are grouped into K sub-blocks of N/K symbols. The case of blocks with contiguous carriers has the advantage of simplicity and it is more suitable for detection systems. The case of non-contiguous carrier blocks offers better peak factor (PF) reduction capability at the cost of extra complexity. The method generates a set of signal derivates by rotating the symbols from each block with one phase from a given set of K phases with values from a given finite set. Then, after calculation of the corresponding PAPR of each signal variant, the one with minimal PAPR is chosen for the transmission.

Both methods provide efficient PAPR reduction, having the drawback of additional side information required to be sent to receiver. Another disadvantage of these methods is that the complexity of computation is increasing with the number of phase set and block number.
Optimizations of these methods have been proposed in several papers [3][4].

Another PAPR reduction method is tone reservation (TR), which uses a set of reserved set of subcarriers (tones) to generate signals with lower PAPR level. Besides the advantage of no additional distortion, this method also doesn't need to transmit additional information to the receiver. Because not all subcarriers are used to transmit useful information, this method is considered to lower the data rate of the OFDM-based systems.

Since the development of the original tone reservation method, in order to reduce the computation complexity and to improve the performance, several derivate techniques have

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

been proposed: selective mapping of partial tones (SMOPT) [6], One-Tone One-Peak (OTOP) [7] and one-by-one iteration [8].

Another optimized variant of this method proposes to generate tones for the K largest peaks of the signal. The phases of these tones are chosen to be opposite to $\varphi_j+n\pi/2$, where φ_j is the phase of the identified peaks, j=1, 2, …, K and n=0, 1, 2, 3. The procedure is iterated until convergence reaches the expected threshold [5].

The class of non-linear methods is represented by approaches like active constellation extension (ACE), clipping, partial clipping, and signal compression.

The ACE method change the original OFDM signal by modifying amplitude and phase of tones whose base band modulation symbol is an outer point of the constellation. Those outer signaling points of the conventional constellation are dynamically moved toward outside of the original constellation in order to reduce the PAPR level of the transmitted signal. The domain for allowed alternative points is chosen so that the signal processing does not reduce the constellation's minimum-distance but lowers the PAPR level [15][17].

For additional PAPR reduction, some proposed derivate methods consider outers points projection onto squares or circles around all the QAM constellation points and intentional distortion within the allowed bounds. The tradeoff between level of the constellation distortion and PAPR level is analyzed and optimized as well [16].

The clipping method is another well known non-linear PAPR reduction technique, where the amplitude of the signal is limited to a given threshold. Taking in consideration the fact that the signal must be interpolated before A/D conversion, a variety of clipping methods has been proposed. Some methods suggest the clipping before interpolation, having the disadvantage of the peaks regrowth. Other methods suggest the clipping after interpolation, having the disadvantage of out-of-band power production. In order to overcome this problem different filtering techniques have been proposed. Filtering can also cause peak regrowth, but less than the clipping before interpolation [9].

Another clipping technique supposes that only subcarriers having the highest phase difference between the original signal and its clipped variant will be changed. This is the case of the partial clipping (PC) method [10].

For additional PAPR reduction, some papers proposed μ-law/A-law companding functions [13], exponential companding function [12], piecewise-scales [11] or polynomial ratio functions [14] after the clipping.

The rest of the paper is organized as follows. The second section describes the OFDM signal, some of its properties, and some aspects of the high power amplifier. The third section describes the proposed hybrid PAPR reduction scheme. In the fourth section is described the clipping method as PAPR reduction method of reference. Next, the numerical results highlighted by the computer simulation are presented and discussed in the fifth section. Based on the obtained results, some conclusions are presented in the sixth section.

II. THE OFDM SIGNAL

In an OFDM-based system, the signal samples are grouped in blocks of N symbols, $\{X_n, n=0,1,\ldots,N-1\}$, which are modulating a set of N subcarriers, with frequencies $\{f_n, n=0,1,\ldots,N-1\}$. These subcarriers are chosen to be orthogonal, that is $f_n=n\Delta f$, where $\Delta f=1/T$, and T is the OFDM symbol period. The resulting signal can be written as:

$$x(t)=\frac{1}{\sqrt{N}}\sum_{n=0}^{N-1}X_n e^{j2\pi f_n t}. \qquad (1)$$

In order to avoid the intersymbol interference (ISI) generated by the multipath channels, each signal period is extended with a fraction from itself, corresponding to a guard interval. After Digital-to-Analogue (D/A) conversion, the signal is applied to the modulator. Next, the resulted signal is applied to a high-power amplifier (HPA) which drives the antenna load.

At the receiver, after demodulation, the guard will be removed, the symbols being evaluated for a time interval of $[0,T]$.

Time domain samples of the low-pass OFDM signals in the complex domain are appreciatively Gaussian distributed due to statistical independence of subcarriers. Due to this fact, sporadically, the signal presents peaks, which cause the PAPR problem. The expression of the PAPR for a given OFDM signal block is given by:

$$PAPR(x)=\frac{\max\left(|x(t)|^2\right)}{E\left[|x(t)|^2\right]}, \qquad (2)$$

where $E[.]$ denotes the expectation operator.
This is usually evaluated using the complementary cumulative distribution function (CCDF) of the PAPR:

$$\begin{aligned} CCDF(Y)&=\Pr(PAPR>Y)=\\ &=1-\Pr(PAPR<Y)\end{aligned} \qquad (3)$$

where Y is a PAPR threshold.

Another quality measure refers to the non-linearity of the transmitted signal which is produced by the HPA. This is the Signal-to-Distortion Ratio (SDR) defined as:

$$SDR=\frac{\|x\|^2}{\|x-g(x)\|^2}, \qquad (4)$$

where $g(.)$ is the memoryless nonlinearity representing the effects of the HPA.

In order to describe these effects, several models have been proposed. One of the well known models is the Saleh Model, which is described by the following input-output equations:

$$A_{HPA}(u) = \frac{\alpha \cdot u}{1 + \beta \cdot u^2}, \quad (5)$$

$$P_{HPA}(u) = \frac{\alpha \cdot u^2}{1 + \beta \cdot u^2}, \quad (6)$$

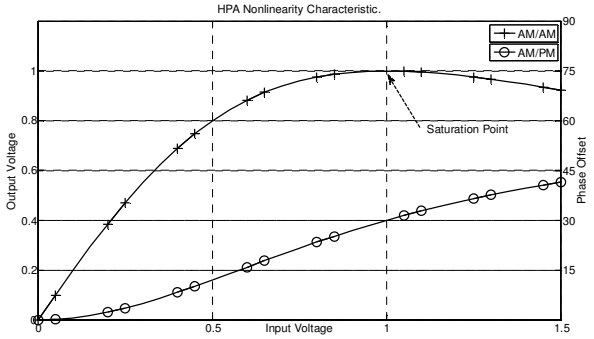


Figure 1. Example of a HPA nonlinear characteristic. Amplitude and phase transfer functions are presented.

where $A(u)$ is the amplitude characteristic and $P(u)$ is the phase characteristic.
An example for the coefficients with the values of α=2 and β=1, is presented in Figure 1.

The optimal solution for PAPR problem may not be the best solution for the SDR problem and vice versa. Because these two problems are correlated, in practice a suboptimal solution may be chosen [15].

III. THE HYBRID METHOD

The proposed hybrid PAPR reduction technique is obtained by serialization of active constellation extension method and sequential tone reservation method.

The main idea for combining the two methods is relying on the observation that the cumulative signal processing for PAPR reduction will increase the overall performance. Furthermore, the idea is based on the fact that each of the considered methods is based on a different principle. One performs a controlled signal distortion and the other realizes different changes of the non-data subcarriers.

The block diagram of the proposed method is presented in Figure 2. The performance of the proposed PAPR reduction technique is analyzed with a MATLAB simulator as presented in Figure 3. Within this simulator, the samples from the generated signal are mapped from binary representation to the M-QAM or M-PSK constellation points. The obtained complex values are grouped in blocks of N elements each, forming the OFDM symbols.
The obtained OFDM frames are applied to the PAPR reduction blocks. For a better performance comparison, besides the proposed ACE-TR method, additionally the clipping method [9] is taken into consideration.

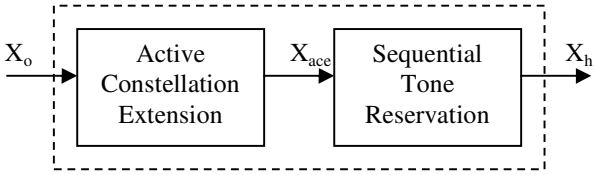


Figure 2. The Hybrid ACE-TR scheme for PAPR reduction.

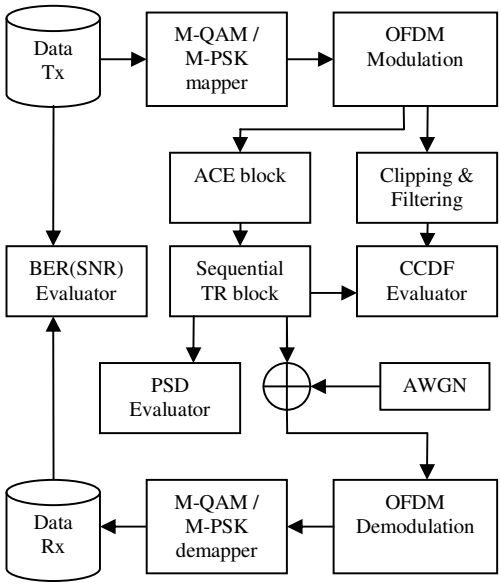


Figure 3. MATLAB model for the analysis of the hybrid PAPR reduction technique.

The PAPR reduction blocks alter the original signal. Due to this fact, for evaluation of communication's performance and efficiency, the simulator estimates the bit error rate (BER) and power spectral density (PSD) for the signal obtained after processing for PAPR reduction.

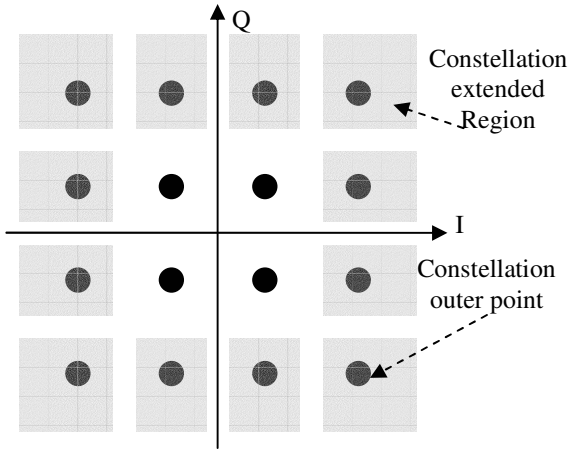


Figure 4. Example of an extended constellation. The original points and allowed extended domain for outer points for 16-QAM are indicated.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

The ACE method requires both time-domain and frequency-domain signal processing. As already mentioned, the main idea of this method is to shift the outer constellation points toward exterior of original constellation generating an alternative representation of the same symbol. The allowed domains for these outer points are presented in Figure 4, when 16-QAM is used as base band modulation.

The boundaries of these domains are constrained by the constellation's minimum-distance. The reason for this limitation is to prevent a decrease of BER performance at the receiver.

In the present work, the ACE was implemented according with the following algorithm [15]:

1) Starting with a block of N data symbols X_o, representing a frequency domain OFDM frame;

2) Apply IFFT to get the corresponding time-domain signal representation $x[n]$.

3) Clip any sample which satisfies the condition: $|x[n]| \geq A$, to obtain a signal with reduced maximal amplitude modulus:

$$\tilde{x}[n] = \begin{cases} x[n], & |x[n]| \leq A \\ A \cdot e^{j\theta[n]}, & |x[n]| \geq A \end{cases}, \quad (7)$$

where

$$x[n] = |x[n]| \cdot e^{j\theta[n]}, \quad (8)$$

4) Apply FFT to obtain the frequency-domain representation $\tilde{X}$ of the previous time-domain clipped signal $\tilde{x}$;

5) Enforce all ACE constraints on $\tilde{X}$ by restoring the original values for all inner points and limit the values for all outer points to reside within their corresponding extended domains as already mentioned;

6) Return to step 2 and iterate the algorithm until the maximum number of iterations is reached or the PAPR level decreased until a given threshold.

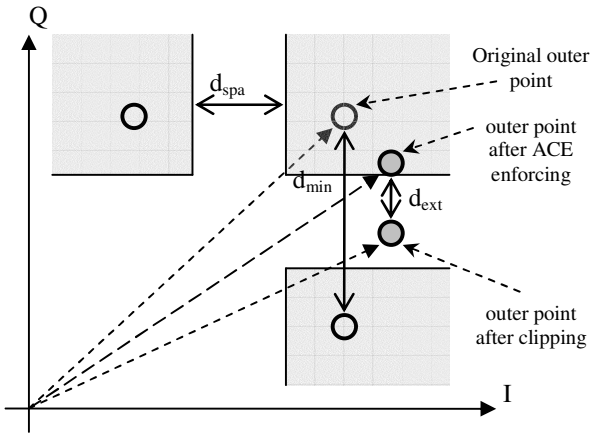


Figure 5. ACE outer point enforcing. Exemplification for 16-QAM.

In order to apply the ACE constraints for the outer points of the constellation, their corresponding vectors must be changed. Depending on their relative location versus the location of the original point and on the allowed extended domain, the algorithm may change vector's amplitude or phase or both.

A solution would be to rotate the vector until the outer point enters back into corresponding extended domain. When this operation is not enough, the vector's amplitude can be increased accordingly.

In the present work, in order to reduce the computation complexity, the ACE constraints are applied by checking and changing the Cartesian coordinates separately. The present algorithm considers the constellation's minimum-distance d_{min}, the distance between two adjacent extended domains d_{spa}, the coordinate of the original outer point, and the coordinate of the actual point obtained after clipping. When the in-phase or quadrature value is under the threshold indicated by the corresponding domain border, the difference d_{ext} is added to shift the point back into its domain. This approach is presented in Figure 5. For the case of the M-PSK modulation, the ACE constraint is applied using a similar approach. The difference is that for the M-PSK case, all constellation points can be used to lower the PAPR, and their corresponding extended domains are represented by radial sectors with the angular size of $2\pi/M$ radians.

Next, the obtained frequency-domain signal is applied to the tone reservation method, which is presented in Figure 6. It selects T pilot tones positions from a complete set of Q no-data carrier positions and a set of M complex values, forming a set of M^T possible combinations.

This search space may lead to an increased amount of data computation. The chosen tone reservation algorithm decreases the computation complexity by attempting a reduced search space by trying all M values on the first pilot P[0], while the other pilots, P[1],…,P[T-1], have a "randomized" or zero initial state. Once an optimal value is found, a similar procedure is repeated on the other pilot positions. For further computation complexity reduction, the time-domain signals equivalent for all pilot tones can be computed and stored initially into memory. In this case, more operations are done in time-domain, fact which determines a decreased number of FFT operations [5].

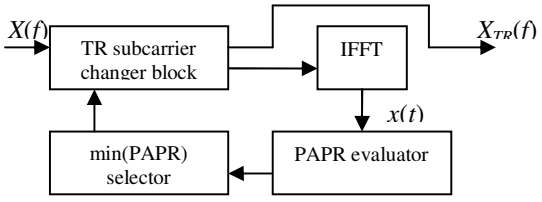


Figure 6. Sequential Tone Reservation method.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

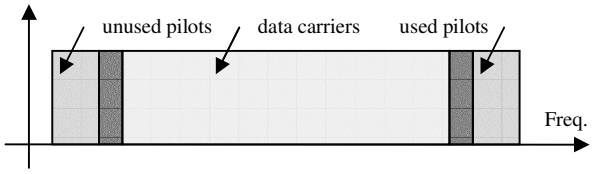


Figure 7. Allocation of reserved tones within an OFDM symbol. (symmetrical) – Type I

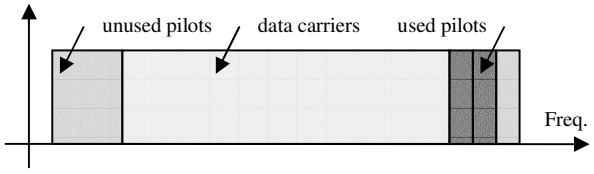


Figure 8. Allocation of reserved tones within an OFDM symbol (lateral, inner) – Type II

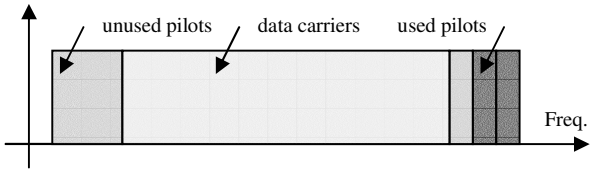


Figure 9. Allocation of reserved tones within an OFDM symbol (lateral, outer) – Type III

Because the TR method operates on some subcarriers from the frequency-domain signal, the displacement of these non-data subcarriers may impact the method's performance. Depending by position, the allocation of the reserved subcarriers may be symmetrical or lateral occupying the lower or higher part of the signal's spectrum. The considered variants are presented in Figures 7 and 8.

When the TR block follows after the ACE block, as Figure 2 indicates, the interfacing of these blocks has to be properly adapted.

Both PAPR reduction blocks have to operate on the same signal. Therefore, in order to have same frequency spectrum, the non-data subcarriers used within TR block has to be available at the ACE block's input.

The ACE block performs a nonlinear signal processing, which will affect the non-data subcarriers as well. Because these subcarriers are not carrying any information, they have no ACE constraints as method requires for the constellation points of the data subcarriers. From the ACE perspective these non-data subcarriers have the optimal value for lowering PAPR.

Contrary, the TR block will change the values of these non-data subcarriers in order to search an OFDM alternative signal with a decreased PAPR. Because some of these subcarriers will provide no improvement from the PAPR reduction point of view, the initial value set of the ACE block has to be considered.

In order to make the proper adaptation, the TR constellation point set of each non-data subcarrier has to include the initial value obtained after previous signal processing performed by the ACE block.

IV. THE CLIPPING METHOD

For analysis of the efficiency of the proposed hybrid technique, also a pure nonlinear method has been considered. This is the clipping method with frequency-domain filtering as presented in [9]. The block diagram of this method is presented in Figure 10. It consists in a zero padding block, an IFFT block, an effective clipping block, and a frequency domain block. For a frequency-domain input signal represented by a vector A_{in} with N elements $[a_0,\ldots,a_{N-1}]$ and an oversampling factor p, the zero padding inserts $N(p-1)$ zeros in the middle of this vector, forming the new vector A_{zp}. The clipping block limits the amplitude of the time-domain signal to a given threshold. The resulted signal a_{clp} is then applied to the frequency domain filter where the output signal a_{out} is obtained.

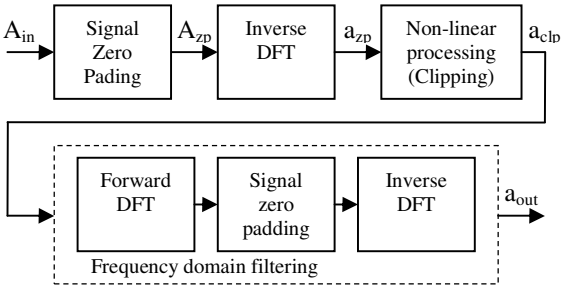


Figure 10. Clipping with filtering PAPR reduction method.

The clipping ratio (*CR*) applied in this method is defined as ratio of the clipping level *A* to the root-mean-square power σ of the unclipped baseband signal,

$$CR = 20 \cdot \log_{10}\left(\frac{A}{\sigma}\right). \quad (9)$$

The filtering block is composed by an FFT block, another zero padding block and an IFFT block. It is designated to reduce the out-of-band noise without distorting the in-band discrete signal.

V. NUMERICAL RESULTS

The MATLAB simulations have been performed for base-band signals with N=128 subcarriers using M-QAM and M-PSK modulations. The frequency-domain signal is extended with additional Q=24 no-data subcarriers. From this set, T=12 subcarriers are used for PAPR reduction by the TR method. The corresponding constellation consists in sets of M=16 points. For the reference clipping method, the simulation considers the clipping rate *CR* having some values in the range of 6-14 and the oversampling factor p set to 2.

For the OFDM signal spectrum computation, it was considered that the distance between two adjacent subcarriers is 0.2 MHz.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

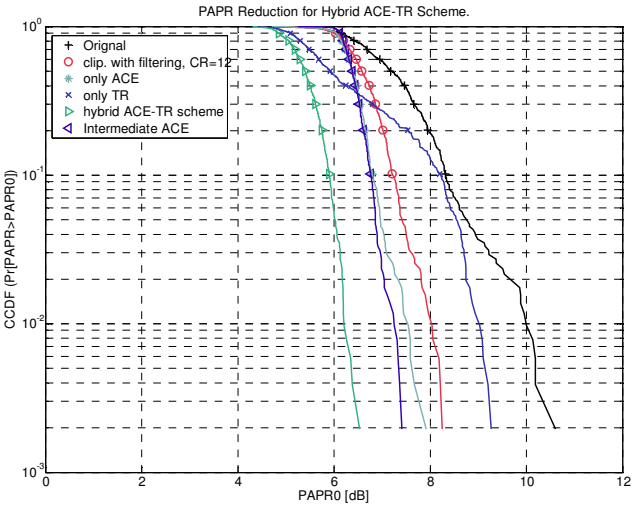


Figure 11. PAPR reduction using hybrid ACE-TR method. Type I

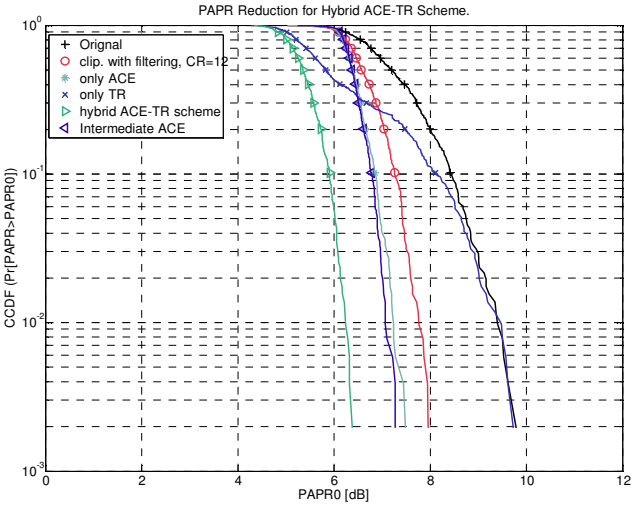


Figure 12. PAPR reduction using hybrid ACE-TR method. Type II

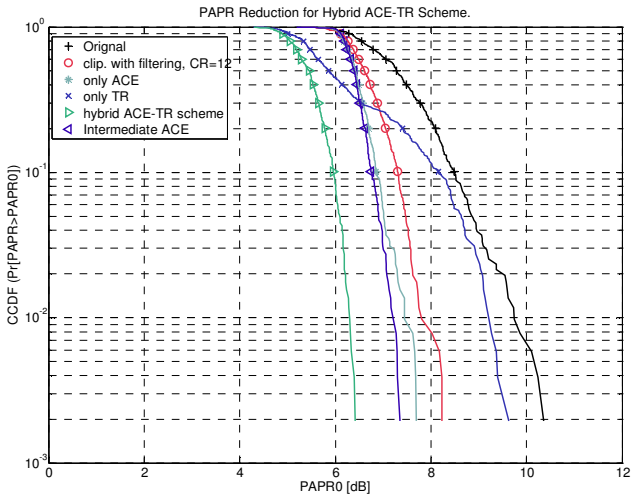


Figure 13. PAPR reduction using hybrid ACE-TR method. Type III

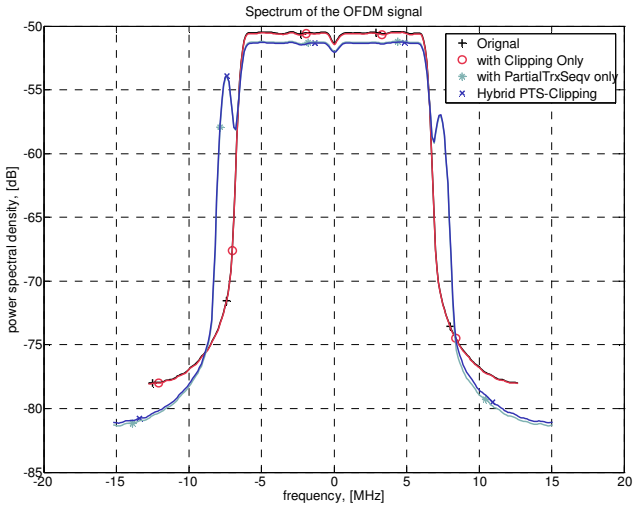


Figure 14. Spectr. of OFDM signal before/after PAPR reduction. Type I

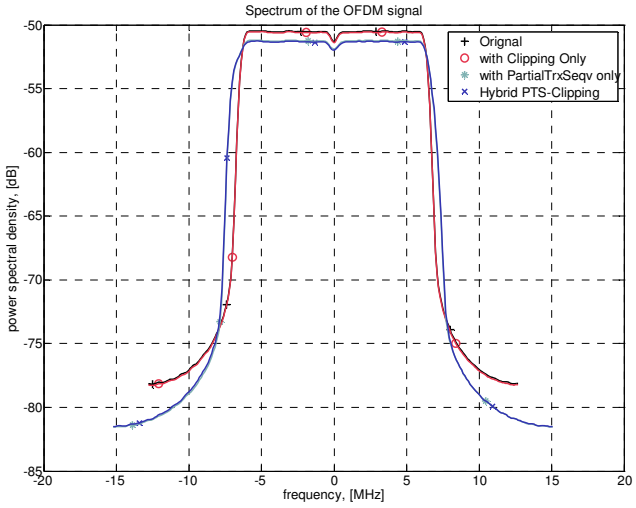


Figure 15. Spectr. of OFDM signal before/after PAPR reduction. Type II

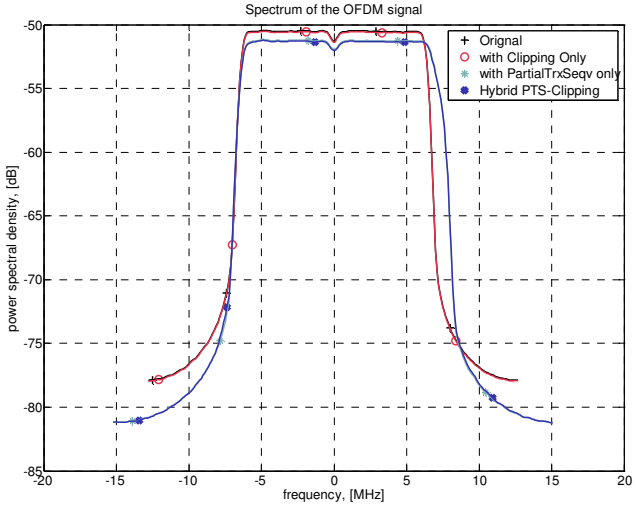


Figure 16. Spectr. of OFDM signal before/after PAPR reduction. Type III

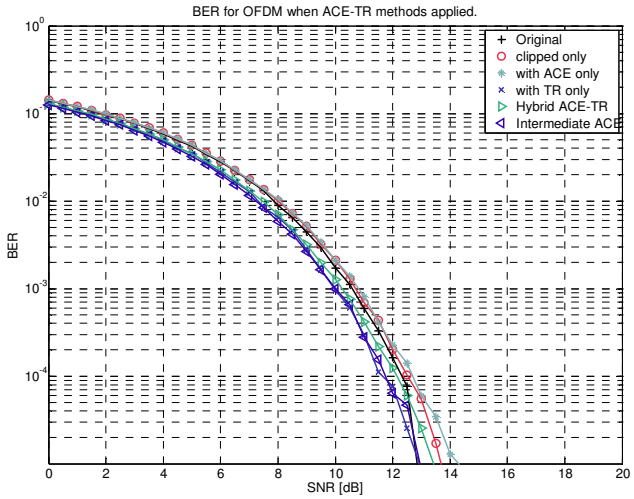


Figure 17. BER of OFDM signal before and after PAPR reduction.

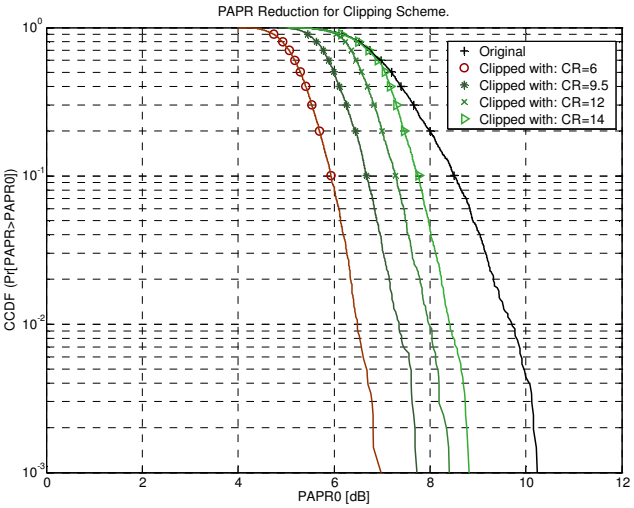


Figure 18. PAPR reduction using clipping method.

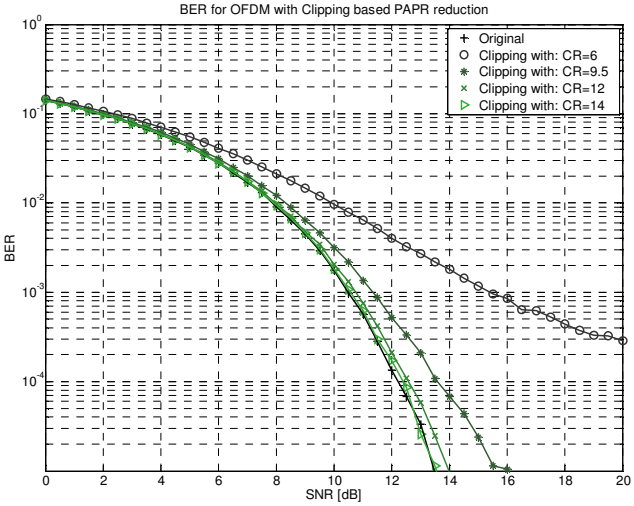


Figure 19. BER of OFDM signal befor/after clipping.

The results presented in this paper are obtained for OFDM frames with the repartition of non-data subcarriers as previously indicated in Figures 7, 8, and 9, with constellations of the pilot search space points identically with the constellations of the constellation used for data carriers.

The numerical results show that the proposed scheme improves the PAPR reduction in comparison with the use of only one of the component methods. This improvement is highlighted in this section with three cases.

In the first case, Figures 11 and 14 indicate the PAPR reduction and signal's frequency spectrum when the OFDM frame has the non-data subcarriers configuration as presented in Figure 7.

In Figure 11, it can be observed that the ACE method performs a better PAPR reduction while the applied TR method obtains a lower PAPR reduction than the clipping at a ratio of *CR*=12. A slight difference can be observed between the ACE applied on the initial OFDM frame and the extended OFDM frame containing the reserved non-data subcarriers. The hybrid ACE-TR provides better PAPR reduction since it accumulates the effects from the two methods.

Due to the insertion of the additional non-data subcarriers on the both sides of the original spectrum, the obtained signal presents an increased bandwidth as indicated in Figure 14.

In the second case, Figures 12 and 15 indicate the PAPR reduction and signal's frequency spectrum for the configuration shown in Figure 8.

Due to a different displacement of the non-data subcarriers; the TR method has a different efficiency for the PAPR reduction. Even if this method has smaller PAPR reduction, with the hybrid method still higher PAPR reduction is obtained. Also, this case presents a smaller increase of the bandwidth for the resulted OFDM signal than the one from the previous one.

In the third case, similarly, Figures 13 and 16 indicate the same signal parameters when the non-data subcarriers are located as is shown in Figure 9.

For the PAPR reduction, this case is quite similar with the first one. The difference consists on the spectrum of the resulted signal, which has a slightly asymmetrical shape.

Figure 17 shows that the BER performance is slightly influenced by the proposed PAPR reduction technique, being better than in case of simple clipping.

For a better evaluation of the performance of the proposed method, the PAPR reduction and corresponding BER characteristic of the clipping method are presented in Figure 18 and Figure 19, respectively.

The simulations shown that, if smaller values for the clipping ratio are considered, the clipping method obtains comparable PAPR reduction as the hybrid method do.

The drawback of this case is that the smaller *CR* values imply an increased signal distortion, and so a worst BER performance.

Therefore the presented numerical results shown that, in all cases, the hybrid ACE-TR method provides better PAPR

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

reduction than in case of use of only one component method. Additionally, compared with clipping, the combined technique presents no degradation of the BER performance.

The computational complexity of the algorithm of the hybrid technique is given by the sum of the computational complexity of the component methods.

The ACE block performs one IFFT, one clip, one FFT and one vector shift per iteration. For the present simulations, we have limited the number of iterations in the ACE block to one, therefore the amount of operations for this block is $O(2+2\cdot N\cdot \log_2(N))$.

The TR block performs one change for a pilot subcarrier and one IFFT per iteration. Considering the applied algorithm and the size of the search space, a complete operation requires $O(M\cdot T\cdot(1+N\cdot \log_2(N)))$.

In practice, the amount of operations can be reduced if the time-domain signal corresponding to each single non-data subcarrier is pre-computed and stored in a nonvolatile memory. In this case the amount of operations is reduced to $O(M\cdot T+N\cdot \log_2(N))$.

Based on these expressions, it can be observed that the amount of operations required by the hybrid method is bigger than the number of operations required by other PAPR reduction techniques and depends by number of data subcarriers and the size of the search space used by the TR block.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new PAPR reduction technique based on the combination of an active constellation extension method with a tone reservation method.

The paper presents the ACE and TR algorithms used within the hybrid technique. The interfacing of the two PAPR reduction blocks is also explained.

The simulation results show that the hybrid scheme realizes higher PAPR reduction for various OFDM frame formats. Similar results for PAPR reduction have been obtained for the case when TR block precedes the ACE block.

The two methods considered have various derivates, bringing different efficiency and performance. The ACE method may be implemented using different constellation restrictions, obtaining different PAPR reduction levels and BER performances. The TR method may use different set of values for the non-data subcarriers. Depending on this set, its computation complexity and PAPR reduction strength may significantly vary.

In future work, we will consider different ACE constraints and TR schemes with different numbers and various sets of values for the non-data carriers.

REFERENCES

[1] S. Muller, R. Bauml, R. Fischer, and J. Huber, “OFDM with reduced peak-to-average power ratio by multiple signal representation”, Annals of Telecommunications, vol. 53, pp. 58–67, February 1997.

[2] L. J. Cimini Jr. and N. R. Sollenberger, “Peak-to-average power ratio reduction of an OFDM signal using partial transmit sequences”, IEEE Commun. Lett., vol. 4, no. 3, pp. 86–88, March 2000.

[3] L. Wang and Y. Cao, “Improved SLM for PAPR Reduction in OFDM Systems”, International Workshop on Intelligent Systems and Applications, pp. 1-4, May 2009.

[4] L. Wang and J. Liu, “PAPR Reduction of OFDM Signals by PTS With Grouping and Recursive Phase Weighting Methods”, IEEE Transactions on Broadcasting, vol. 57, no. 2, pp. 299-306, June 2011.

[5] Y.Z. Jiao, X.J. Liu, and X.A. Wang, "A Novel Tone Reservation Scheme with Fast Convergence for PAPR Reduction in OFDM Systems", Consumer Communications and Networking Conference, pp. 398-402, January 2008.

[6] S. Yoo, S. Yoon, S.Y. Kim, and L. Song, “A novel PAPR reduction scheme for OFDM systems: selective mapping of partial tones (SMOPT)”, IEEE Transaction on Consumer Electronics, vol. 52, no. 1, pp. 40-43, February 2006.

[7] E. Bouquet, S. Haese, M. Drissi, C. Moullec, and K. Sayegrih, “An innovative and low complexity PAPR reduction technique for multicarrier systems,” Proc. 9th European Conference on Wireless Technology, pp. 162-165, September 2006.

[8] C. L. Wang, Y. Ouyang, and H. C. Chen, “A low-complexity peak-to-average power ratio reduction technique for OFDM-based systems”, Proc. 60th IEEE Vehicular Technology Conference, vol. 6, pp. 4380-4384, September 2004.

[9] J. Armstrong, “New OFDM Peak-to-Average Power Reduction Scheme”, Proc. of IEEE Vehicular Technology, pp. 756-760, May 2001.

[10] M. Deumal, C. Vilella, J. L. Pijoan, and P. Bergadà, “Partially Clipping (PC) Method for the Peak-to-Average Power Ratio (PAPR) Reduction in OFDM”, IEEE Int. Symposium on Personal, Indoor and Mobile Radio Communications, vol. 1, pp. 464-468, September 2004.

[11] X. Huang, J. Lu, J. Zheng, and J. Gu, "Piecewise-scales transform for the reduction of PAPR of OFDM signals", IEEE Global Telecommunications Conference, vol. 1, pp. 564-568, November 2002.

[12] T. Moazzeni, H. Selvaraj, and Y. Jiang, "A Novel Multi-Exponential Function-based Companding Technique for Uniform Signal Compression over Channels with Limited Dynamic Range", Intl. Journal of Electronics and Telecommunications, Vol. 56, No. 2, pp. 125-128, June 2010.

[13] X. Wang, T. T. Tjhung, C. S. Ng, and A. A. Kassim, “On the SER Analysis of A-Law Companded OFDM System”, in Proc. Global Telecommunication Conference, vol. 2, pp. 756-760, December 2000.

[14] V. Cuteanu, A. Isar, “PAPR reduction of OFDM signals using hybrid clipping-companding scheme with sigmoid functions”, International Conference on Applied Electronics, pp. 75-78, September 2011.

[15] B.S. Krongold and D. L. Jones, "PAR Reduction in OFDM via Active Constellation Extension", IEEE Transactions on Broadcasting, Vol. 49, No. 3, pp. 258-268, September 2003.

[16] M. Malkin, B. Krongold, and J. M. Cioffi, “Optimal constellation distortion for PAR reduction in OFDM systems”, IEEE 19th Int. Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1-5, September 2008.

[17] A. Kliks and H. Bogucka, “Improving Effectiveness of the Active Constellation Extension Method for PAPR Reduction in Generalized Multicarrier Signals”, SpringerLink, Wireless Personal Communication, vol. 61, no. 2, pp. 323-334, May 2010.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

PAPR Reduction of OFDM Signals using Partial Transmit Sequence and Clipping Hybrid Scheme

Eugen-Victor Cuteanu
Communication Department
Politehnica University, Faculty of Electronics and Telecommunications
Timisoara, Romania
victor.cuteanu@gmail.com

Alexandru Isar
Communication Department
Politehnica University, Faculty of Electronics and Telecommunications
Timisoara, Romania
alexandru.isar@etc.upt.ro

***Abstract*—The Orthogonal Frequency Division Multiplexing is one of the widely used modulation techniques in the broadband wireless technology. One of the main problems is the high peak-to-average power ratio of transmitted signal due to the superposition of many subcarriers. This paper presents a new hybrid peak-to-average power ratio reduction technique, which combines a partial transmit sequence method with the clipping method. The paper highlights the performance and advantages of the mixed technique and compares it with other existing methods. The simulations shown that the proposed technique realizes an increased peak-to-average power ratio with a decreased signal distortion compared to clipping method.**

Keywords-OFDM; PAPR; Partial Transmit Sequence; Clipping

I. INTRODUCTION

The Orthogonal Frequency Division Multiplexing (OFDM) is used in broadband wireless communication systems like Worldwide Interoperability for Microwave Access (WiMAX), Terrestrial Digital Video Broadcast (DVB-T), or wireline systems like Asymmetric Digital Subscriber Line (ADSL). The main problem of the OFDM is the high value of Peak-to-Average Power Ratio (PAPR) of the transmitted signal. Due to the superposition of the many data subcarriers, the OFDM signal exhibits Rayleigh-like characteristics. The large amplitude variations lead to high values for the PAPR. These peaks require the high power amplifiers (HPA) to support wide linear dynamic range.

Higher signal level at the input of HPA causes non-linear distortions at its output, leading to an inefficient operation of HPA. These distortions cause intermodulation products resulting unwanted out-of-band power. In order to reduce the PAPR of OFDM signals, many solutions have been proposed and analyzed. Some of the main characteristics of these methods are non-linearity, computation complexity and size of side information needed to be sent to receiver.

Some of the well known linear methods are selective mapping (SLM) [1], partial transmit sequence (PTS) [2], and tone reservation (TR) [5].

In the SLM method, the vectors from the original frequency domain OFDM signal are rotated based on a set of predefined phase arrays. For each signal variant obtained, its corresponding PAPR is evaluated. The one with the lowest PAPR is chosen for the transmission.

The PTS method uses a similar principle with the difference that same rotation angle is applied to more than one vector. The method considers the N complex values representing OFDM signal vectors as being grouped into K sub-blocks of N/K elements each. The case of blocks with contiguous carriers has the advantage of simplicity and is more suitable for detection systems. The case of non-contiguous carrier blocks offers better peak factor (PF) reduction capability at the cost of extra complexity.
The method generates a set of signal variants by rotating the vectors from each block with one phase from a given set of K phases with values from a given finite set. Then, after calculation of the corresponding PAPR of each signal variant, the one with minimal PAPR is being chosen for the transmission.

The efficiency of these methods increases with the number of phases from the considered set. The efficiency of the PTS method also increases when a higher number of blocks are used. The disadvantage is that a better efficiency requires an increased amount of computation at the transmitter's side and receiver's side. Because the receiver must know those phases' sets and block sizes, another drawback of these methods is the additional information required to be sent to receiver. Optimizations of those methods have been proposed in several papers [3, 4].

The TR method represents another linear technique which instead of altering the existing data subcarriers modifies the vectors from an additional set of non-data subcarriers. The method calculates the signal's PAPR values which correspond to the allocated reserved subcarriers. The signal replica corresponding to minimal PAPR is chosen for the transmission.

The important advantage of this technique is that at receiver's side no additional information and no computation are required. Because not all subcarriers are used to transmit useful information, this method has the disadvantage of a lower data rate.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

In order to reduce the computation complexity and to improve the performance, several PAPR reduction techniques were derived from the original tone reservation method: selective mapping of partial tones (SMOPT) [6], One-Tone One-Peak (OTOP) [7] and one-by-one iteration [8], fast TR described in [5].

The class of non-linear methods is represented by approaches like clipping, partial clipping (PC), signal compression and active constellation extension (ACE).

The clipping method is another very well known non-linear PAPR reduction technique, where the amplitude of the signal is limited to a given threshold.

Taking in consideration the fact that the signal must be interpolated before A/D conversion, a variety of clipping methods has been proposed. Some methods suggest the clipping before interpolation, having the disadvantage of the peaks regrowth. Other methods suggest the clipping after interpolation, having the disadvantage of out-of-band power production. In order to overcome this problem, different filtering techniques have been proposed. Filtering can also cause peak regrowth, but less than the clipping before interpolation [9].

Signal compression is another group of non-linear methods which improves the PAPR reduction. For this purpose some papers proposed μ-law/A-law companding functions [13], exponential companding function [12], piecewise-scales [11] or polynomial ratio functions [14] after the clipping. One of the drawbacks of these methods is the increased noise level generated by the corresponding signal decompression.

The partial clipping is another nonlinear PAPR reduction method which performs additional signal processing in frequency-domain to reduce the distortions of the signal's spectrum. This method supposes that only subcarriers having the highest phase difference between the original signal and its clipped variant will be changed [10].

The ACE method relays on the idea that the outer points from the original constellation may be moved toward outside in order to reduce the PAPR level of the transmitted signal. The domain for allowed alternative points is chosen so that the signal processing does not reduce the constellation's minimum-distance but lowers the PAPR level [15-17].

The rest of the paper is organized as follows. The second section describes the OFDM signal, some of its properties, and some aspects of the high power amplifier. The third section describes the proposed hybrid PAPR reduction scheme. Next, the numerical results highlighted by the computer simulation are presented and discussed. Based on the results obtained, some conclusions are presented.

II. THE OFDM SIGNAL

An OFDM-based communication system uses a complex multicarrier modulation. Each of the N subcarriers, having the frequencies $\{f_n, n=0,1,\ldots,N-1\}$, are modulated with a data sample from a block of symbols $\{X_n, n=0,1,\ldots,N-1\}$.

In order to reduce the bandwidth of the required frequency spectrum, the subcarriers are chosen to be orthogonal, that is $f_n=n\Delta f$, where $\Delta f=1/T$, and T is the OFDM symbol period.

Considering these aspects, the resulting signal can be written as:

$$x(t)=\frac{1}{\sqrt{N}}\sum_{n=0}^{N-1}X_n e^{j2\pi f_n t}\ . \qquad (1)$$

In the many cases, the OFDM-based system communicates through multipath channels, which generate intersymbol interference (ISI). In order to avoid this problem, a fraction from the current signal period, corresponding to a guard interval, is added to the signal. After Digital-to-Analogue (D/A) conversion, the signal is modulated and applied to a HPA. The amplified signal is applied to the antenna.

At the receiver, after demodulation, the guard interval will be removed, the symbols being evaluated for time intervals of length T.

Due to statistical independence of subcarriers, the low-pass time-domain OFDM signal in the complex domain presents a Gaussian distribution. Therefore, sporadically, the signal presents peaks, causing the PAPR problem. The expression of the PAPR for a given OFDM signal block is given by:

$$PAPR(x)=\frac{\max\left(|x(t)|^2\right)}{E\left[|x(t)|^2\right]}, \qquad (2)$$

where E[.] denotes the expectation operator.
The PAPR is usually evaluated using the complementary cumulative distribution function (CCDF) of the PAPR:

$$\begin{aligned} CCDF(Y)&=\Pr(PAPR>Y)=\\ &=1-\Pr(PAPR<Y). \end{aligned} \qquad (3)$$

The non-linearity of the transmitted signal, produced by the HPA, can be evaluated with another quality measure. This is the Signal-to-Distortion Ratio (SDR) defined as:

$$SDR=\frac{\|x\|^2}{\|x-g(x)\|^2}, \qquad (4)$$

where g(.) is the memoryless nonlinearity representing the effects of the HPA.

These effects are described by various models. One of the well known models is the Saleh Model, which is described by the following input-output equations:

$$A_{HPA}(u)=\frac{\alpha\cdot u}{1+\beta\cdot u^2}, \qquad (5)$$

$$P_{HPA}(u)=\frac{\alpha\cdot u^2}{1+\beta\cdot u^2}, \qquad (6)$$

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

where $A(u)$ is the amplitude transfer function, and $P(u)$ is the phase transfer function and α and β are two parameters.
An example for the parameter values α=2 and β=1, is presented in Figure 1.

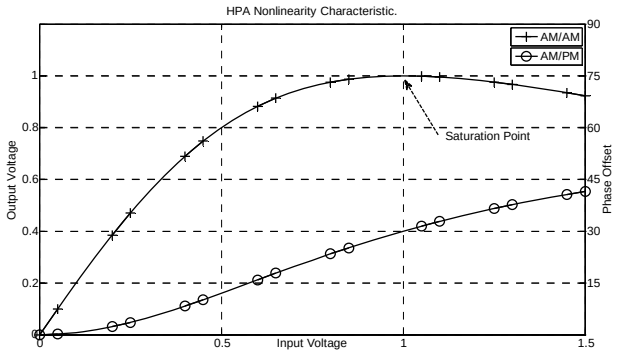


Figure 1. Example of a HPA nonlinear characteristic. Amplitude and phase transfer functions are presented.

The optimal solution for PAPR problem may not be the best solution for the SDR problem and vice versa. Because these two problems are correlated, in practice, a suboptimal solution may be chosen [15].

III. THE HYBRID METHOD

In this section, we present the proposed hybrid PAPR reduction technique which has been obtained by the association of PTS method with clipping method.

The main idea for combining the two methods is relying on the observation that the cumulative signal processing for PAPR reduction significantly improves the overall outcome. Furthermore, the hybrid technique exploits the fact that each of the component methods is based on a different principle.

One performs linear transformation by rotating the vectors from the frequency-domain signal, and the other one performs a non-linear transformation represented by signal limitation. The block diagram of the proposed method is presented in Figure 2.

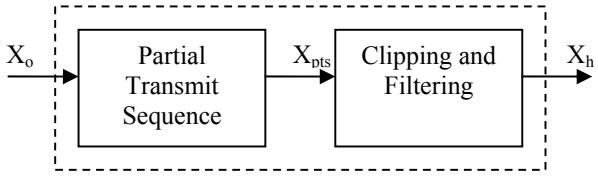


Figure 2. The Hybrid PTS-CLP scheme for PAPR reduction.

The performance of the proposed PAPR reduction technique is analyzed with a MATLAB simulator as presented in Figure 3. Within this simulator, the samples from the generated signal are mapped from binary representation to the M-QAM or M-PSK constellation points. The obtained complex values are grouped in blocks of N elements each, forming the OFDM symbols. The obtained OFDM frames are applied sequentially to PTS block and then to clipping block. For a better evaluation of the proposed method, the results obtained only from clipping [9] are also considered.

The parameters of the resulting signal change according with the signal processing applied by the two PAPR reduction methods.

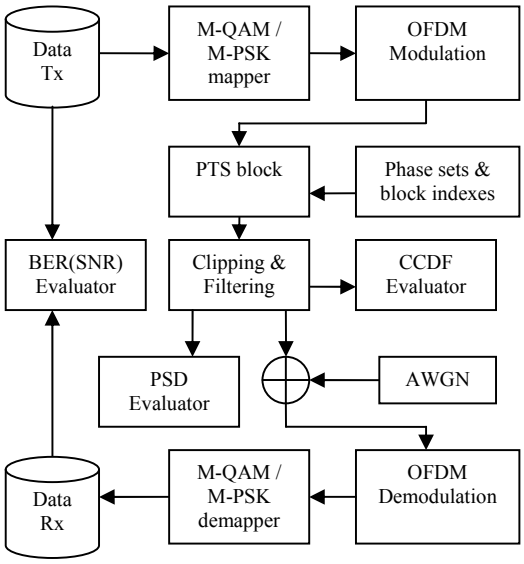


Figure 3. MATLAB model for the analysis of the hybrid PAPR reduction technique.

In order to evaluate the performance and efficiency of o communication system based on the proposed PAPR reduction method, the simulator computes the bit error rate (BER) and power spectral density (PSD) for the original and processed signals.

The PTS method operates on the frequency-domain signal iteratively until the best signal derivate is found. As already mentioned, the main idea of this method is to change the phases of the vectors composing the signal. This method considers the signal's vectors as being grouped in disjoint blocks. The vectors from a block may have a contiguous displacement, or they may be interleaved with the vectors representing another block. The algorithm applies one phase shift for each bloc iteratively until the signal variant having the lowest PAPR is found.

In the present work, the PTS method was implemented considering contiguous blocks of same length each. Additionally, for a better PAPR reduction, the proposed PTS method performs position swap between these blocks. This procedure is depicted in Figure 4.

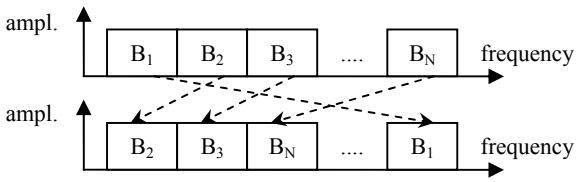


Figure 4. Position swaping between blocks of the OFDM signal.

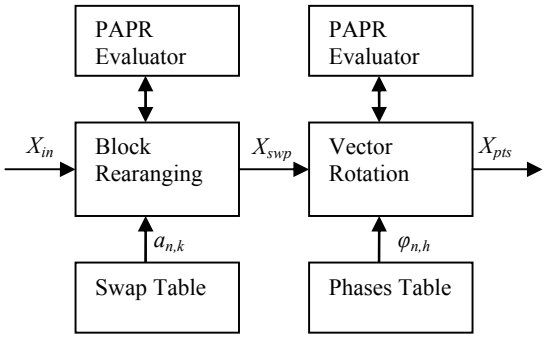


Figure 5. The PTS derivate method for PAPR reduction.

Therefore, the PTS method considered in this paper consists in two stages as the Figure 5 shows.

Based on this block diagram, the PTS algorithm consists in the following steps.

1) Starting with a set of N data symbols X_{in}, representing an OFDM frame:

$$x_{in}(t) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} X_{in}(n) e^{j2\pi f_n t}, \quad (7)$$

2) Rearrange the values X_{in} using the swap table $a_{n,k}$, where n=1...N is the index of a given vector, and k=1…P is the index of a signal variant from a set of P possibilities. Therefore the output signal can be written as:

$$X_{swp}(n,k) = X_{in}(a(n,k)). \quad (8)$$

3) Apply IFFT to get the corresponding time-domain signal representation $x_{swp}[n,k]$.

$$x_{swp}(t,k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} X_{in}(a(n,k)) e^{j2\pi f_n t}. \quad (9)$$

4) Compute the PAPR for all P variants, and choose the one with lowest PAPR level for the following processing step.

5) Rotate all the constellation vectors from X_{swp} using the phase table $\varphi_{n,k}$, where n=1...N is the index of a vector, and h=1…R is the index of the signal variant from a set of R possibilities. The signal becomes:

$$X_{pts}(n,h) = X_{swp}(n) \cdot e^{j\varphi(n,h)}. \quad (10)$$

6) Apply IFFT to get the corresponding time-domain signal representation $x_{pts}[n,h]$.

$$x_{pts}(t,h) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} X_{pts}(n) e^{j(2\pi f_n t + \varphi(n,h))}. \quad (11)$$

7) Compute the PAPR for all R variants, and choose the one with lowest PAPR level for the transmission.

The signal X_{pts} is derived from X_{in} which has modified values for both amplitude and phase of all vectors. Because the receiver must reconstruct the original signal by applying inverse operations, it has to know the block swapping tables and rotation angle tables. For this purpose, it is necessary either to consider predefined tables or to transmit additional information to receiver.

An important drawback of this method is that an increased size of these tables implies increased computation complexity. Furthermore, if the number of block permutations is too high, the possibility for signal reconstruction decreases very much. In such cases, searching of different block position exchanges is not enough, additional information being required.

In order to overcome this problem, we propose a block marking scheme, as presented in Figure 6. The main idea of this model is to use one data carrier as label for each block.

The PTS method changes the phases of all vectors within all blocks. If vector rotation is avoided for the marking subcarriers, PAPR level may not be efficiently reduced. Therefore, the marking model can not relay on the phase information from the additional subcarriers. Additionally amplitude information may not be enough for block indexing, when amplitude levels are less than the number of blocks within the OFDM frame. More, the amplitude is significantly affected by noise.

In order to overcome these problems, the proposed block marking model, uses differential phase information. For this purpose two subcarriers are considered. One represents the phase reference, and the other one represents the block index.

Since amplitudes of these marking vectors do not carry any information, for an improved demodulation, these amplitudes can be increased.

In order to compensate any degradation of the PAPR reduction due to these marking tones, additional two subcarriers are considered. This technique is similar to the tone reservation PAPR reduction method described in [5].

Their phases should be opposite to the phases of the marking subcarriers. Since a tone compensation method is efficient when frequency difference is minimal, their positions are interleaved as in Figure 6.

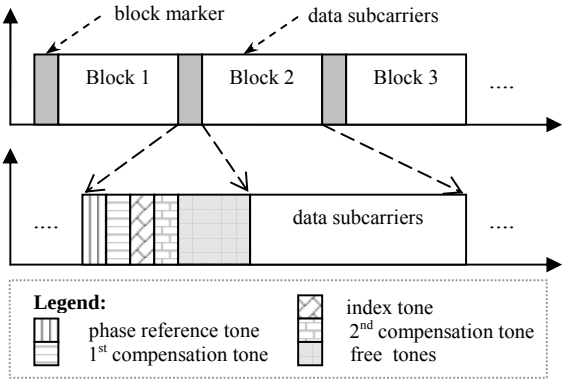


Figure 6. PTS block marking.

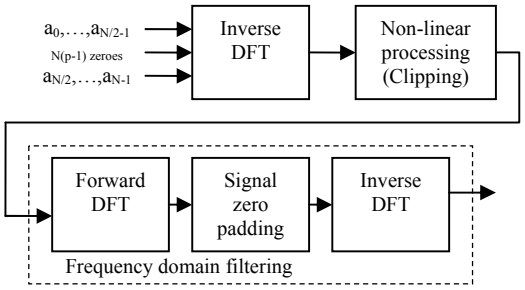


Figure 7. Clipping based PAPR reduction model.

The applied clipping technique [9] is presented in the block diagram from Figure 7.

Here, the input vector $[a_0,\ldots,a_{N-1}]$ is first converted from frequency to time domain using an oversized IFFT. For the oversampling factor p, the input vector is padded with $N(p\text{-}1)$ zeroes placed in the middle of the vector. This results in a trigonometric interpolation of the time domain signal, which fits well for the signals with integral frequencies over original FFT window, like is the case of OFDM. The interpolated signal is then clipped by limiting its amplitude.

The clipping ratio is defined as ratio of the clipping level A and the root-mean-square power σ of the unclipped baseband signal,

$$CR = 20 \cdot \log_{10}\left(\frac{A}{\sigma}\right). \qquad (12)$$

The proposed hybrid PAPR reduction technique contains both linear and non-linear signal processing blocks. This fact produces a high flexibility from the signal processing point of view.

When PTS block is configured to use a reduce set of signal variants, its smaller PAPR reduction can be compensated by the clipping block. This approach presents the advantage of a decreased amount of computation with the price of a higher signal distortion. When PTS block is configured to use an extended set of signal variants, the increased PAPR reduction requires a decreased signal distortion in the clipping block. The drawback in this case is represented by the increased amount of computation.

IV. NUMERICAL RESULTS

The MATLAB simulations have been performed for base-band signals with N=128 and N=256 subcarriers using M-QAM and M-PSK modulations, with corresponding constellations having M=16 points. For the clipping block was considered a clipping rate CR set to 12 and the oversampling factor p set to 2.

For the reference clipping-only method, the simulation considers the clipping rate CR having some values in range of 6-16 and the oversampling factor p set to 2 as well.

For the OFDM signal spectrum computation, it was considered that the distance between two adjacent subcarriers is 0.2 MHz.

The results presented in this paper are obtained using a phase rotation array containing 4 sets of phases with values $\varphi_k = k \cdot \pi / 2$, where k=0…3, and 4 sets of phases with values $\varphi_k = k \cdot 2\pi / 5$, where k=0…4. For the block swapping, the proposed method used a set of 8 randomly chosen combinations for the case when the OFDM signal is composed by 8 blocks of same length. The combinations are chosen in order to present a variety of blocks interleaving. Some of them exchange odd blocks with even blocks, other exchange blocks from first half with blocks from the second half. Also, there are cases when only neighborhood blocks are changed.

The numerical results show that the proposed scheme improves the PAPR reduction in comparison with the use of only one of the component methods. For the evaluation of the performance of the proposed technique, two cases are presented.

The first case considers an OFDM signal with N=128 subcarriers with a 16-QAM modulation. The Figure 8 and Figure 9 indicate the corresponding PAPR reduction and signal's frequency spectrum respectively.

In Figure 8, it can be observed that the modified PTS method with markers and the pure clipping method with CR=12 have similar PAPR reduction. The PAPR reduction is significantly better when the PTS method does not add block markers in the signal. The hybrid PTS-clipping method provides better PAPR reduction since it accumulates the effects from the two methods.

In Figure 9, it can be remarked that in the case of the modified PTS method the signal's spectrum is wider and has a different shape. This is caused by the insertion of additional subcarriers for the block marking and reserved tones. In this case each of the eight blocks contains two free reserved tones.

The second case considers an OFDM signal with N=256 subcarriers with a 16-PSK modulation. The clipping rate is set to same value. The resulted PAPR reduction and signal's frequency spectrum are presented in Figure 11 and Figure 12 respectively.

In Figure 11, it can be observed that the difference between the PAPR reduction of the PTS methods and clipping-only method is even smaller. However, the hybrid method provides better PAPR reduction for this case too.

In Figure 12, it can be remarked that in case of the alternative PTS method, the signal's spectrum is slightly wider, but less than the one from the previous case. This is due to the missing of the free reserved tones.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

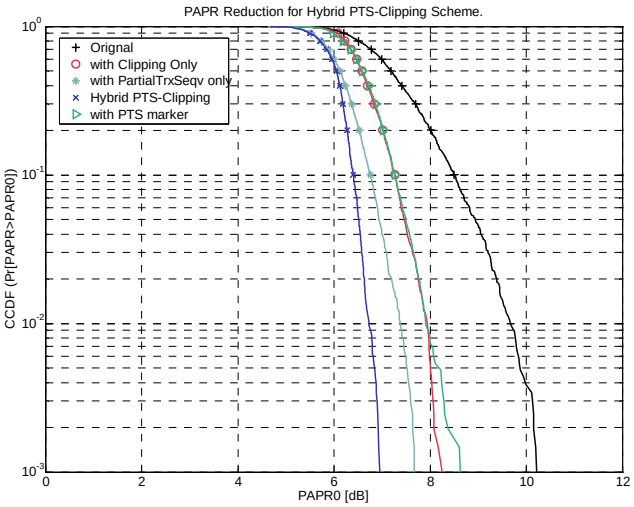


Figure 8. PAPR reduction using hybrid PTS-Clipping method.

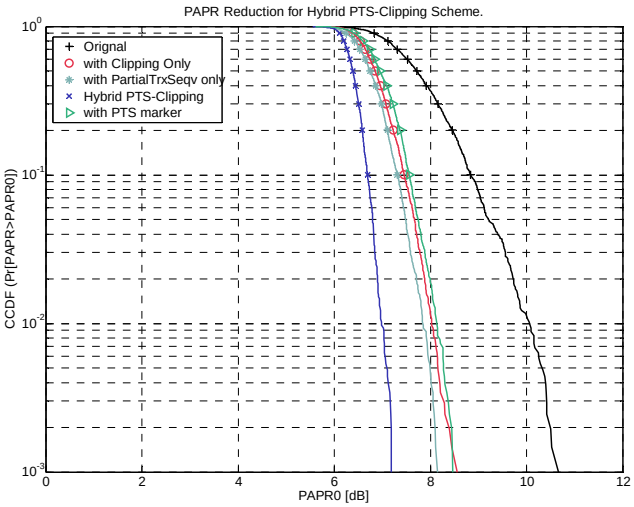


Figure 11. PAPR reduction using hybrid PTS-Clipping method.

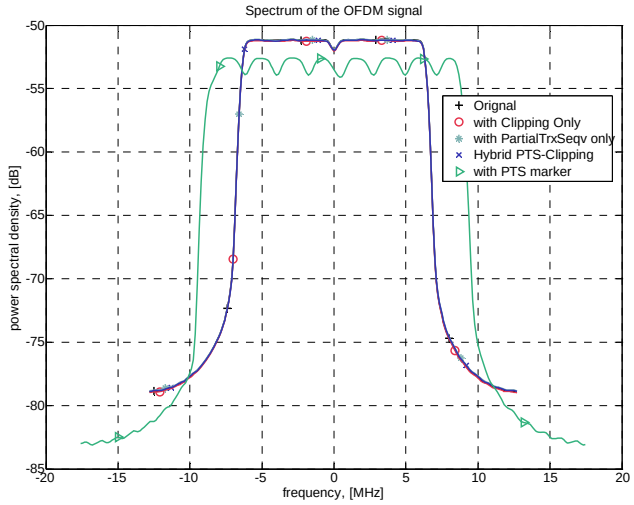


Figure 9. Spectrum of OFDM signal before and after PAPR reduction.

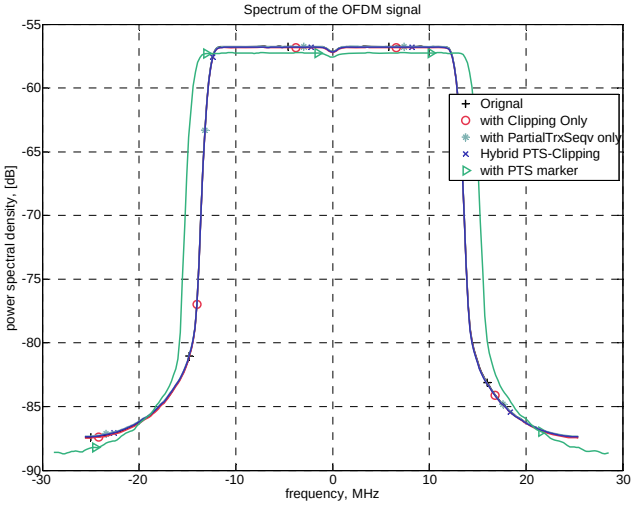


Figure 12. Spectrum of OFDM signal before and after PAPR reduction.

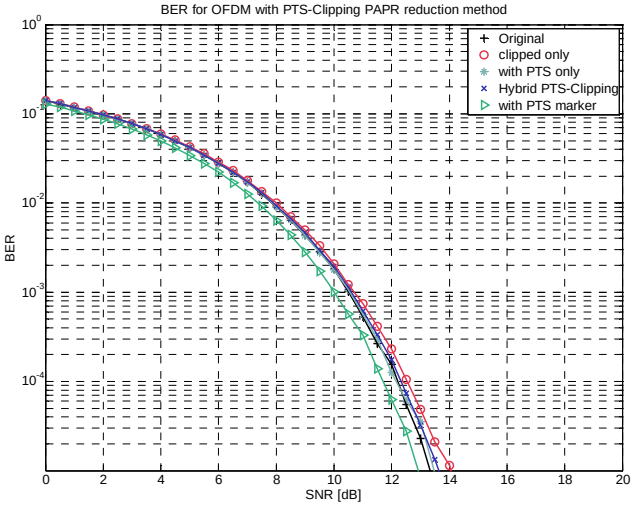


Figure 10. BER of OFDM signal before and after PAPR reduction.

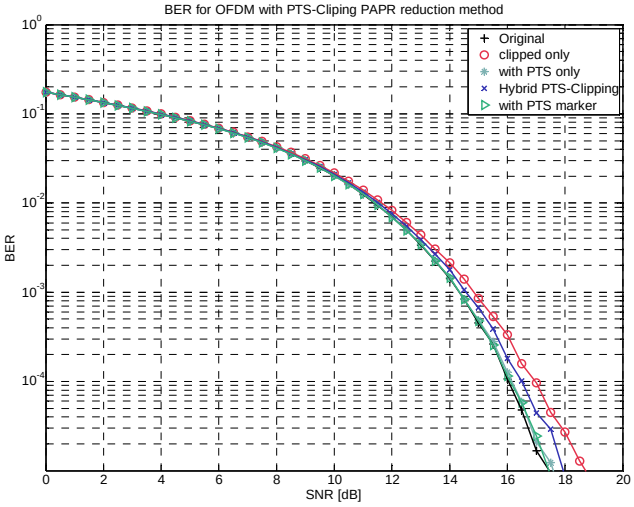


Figure 13. BER of OFDM signal before and after PAPR reduction.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Figure 10 and Figure 13 show that the BER performance is slightly influenced by the proposed PAPR reduction technique, being better than in case of simple clipping.

The PTS method applied in the hybrid scheme does not use the presented block marking technique. The PTS method which includes the block marking technique is separately evaluated.

The Power Spectral Density (PSD) characteristics presented in Figure 9 and Figure 12 highlighted two aspects of the block marking. The first case considers a block marking with two guard subcarriers per block. The second case considers no zero subcarriers within the blocks. The amplitude of the marker subcarriers was set to be equal with the mean value of all data subcarriers.

The required bandwidth for the OFDM signal resulted after the PTS derivate method has increased due of the additional subcarriers. The shape of the frequency spectrum has a slight variation according with the added guard subcarriers.

For the reference clipping method, some PAPR reduction curves together with their corresponding BER characteristics are presented in the Figure 14 and Figure 15, respectively.

The simulations have shown that a smaller value for the clipping ratios increases the PAPR reduction. The case of clipping only with *CR*=6 presents similar results for the PAPR reduction as those obtained with the hybrid method with the previously presented configurations.

The major disadvantage of the clipping only method is that its increased signal distortion leads to a considerable lower BER performance.

Therefore, the presented numerical results shown that, in all cases, the hybrid PTS-clipping method provides better PAPR reduction than in case of use of only one method. Additionally, compared with clipping only method, the combined technique presents insignificantly degradation of the BER performance.

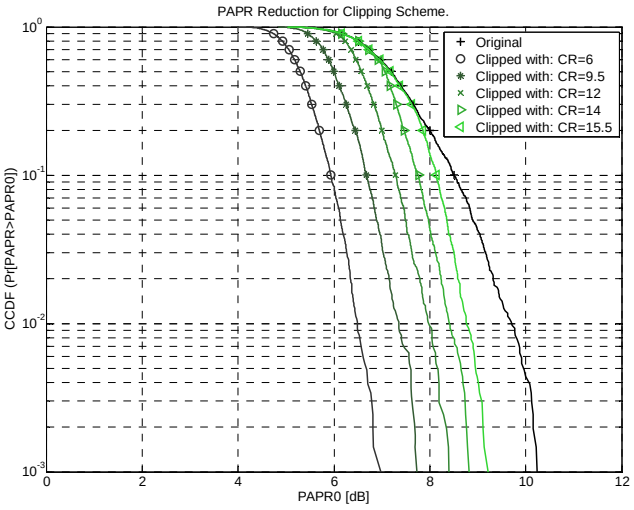


Figure 14. PAPR reduction using clipping method.

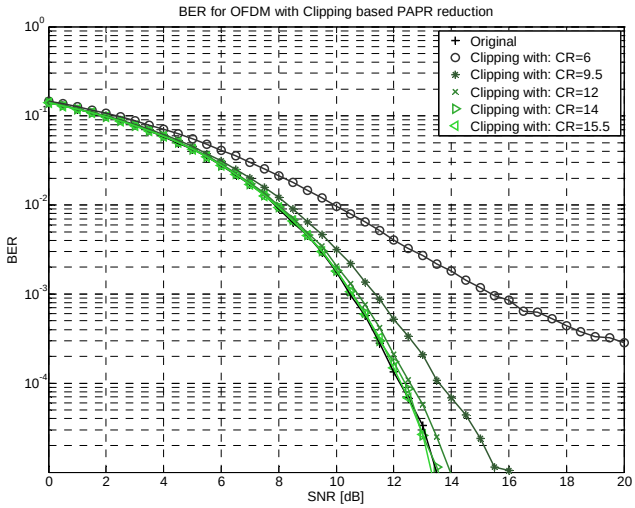


Figure 15. BER of the OFDM signal after clipping.

The computational complexity of the algorithm of the hybrid technique is given by the sum of the computational complexity of the component methods.

The PTS block performs one IFFT, a rearranging of the N subcarriers and change of the phase of all N subcarriers per iteration. Considering the fact that that the PTS method has to perform P block rearranging and R phase changes, the amount of operations corresponding to this component method is $O\left(\left(P+R\right)\cdot\left(1+N\cdot\log_2\left(N\right)\right)\right)$.

For the clipping block, the computation complexity is given by the cumulated amount of complexity for effective clipping and frequency-domain filtering.

The effective clipping block requires one IFFT on the frequency-domain signal with zero padding and $N \cdot p$ amplitude limitations operation on the time-domain signal. Therefore, the amount of operations for this sub-block is $O\left(N\cdot p\cdot\left(1+\log_2\left(N\cdot p\right)\right)\right)$.

The frequency domain filtering performs one FFT, one IFFT and $N \cdot (p-1)$ vector resets. Therefore, the amount of operations corresponding to this sub-block is $O\left(N\cdot\left(p-1\right)+2\cdot N\cdot p\cdot\log_2\left(N\cdot p\right)\right)$.

Finally, the computation complexity for the whole hybrid PAPR reduction method is given by the sum of the three computation complexities.

V. CONCLUSION AND FUTURE WORK

In this paper we proposed an alternative PAPR reduction technique based on the combination of a partial transmit sequence method with the clipping method. The PTS and clipping algorithms used within the hybrid technique are explained.

The numerical results show that the hybrid scheme brings higher PAPR reduction than the component methods.

The considered two methods have few variants with different efficiency and performance.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

The PTS method may be implemented using different block structures, swapping tables and rotation angle tables. Additionally, the block marking can be implemented in various manners, determining different efficiency for PAPR reduction and different sizes of the search space at the receiver's side.

The clipping method, may consider various filtering techniques, each of them leading to various PAPR reduction efficiency.

Therefore the efficiency of the hybrid method may vary when derivates of the component methods are used.

In future work, we will consider different block configurations and different set of values for the tables within the PTS block. Some clipping derivates also may be considered.

REFERENCES

[1] S. Muller, R. Bauml, R. Fischer, and J. Huber, "OFDM with reduced peak-to-average power ratio by multiple signal representation", Annals of Telecommunications, vol. 53, pp. 58–67, February 1997.

[2] L. J. Cimini Jr. and N. R. Sollenberger, "Peak-to-average power ratio reduction of an OFDM signal using partial transmit sequences", IEEE Commun. Lett., vol. 4, no. 3, pp. 86–88, March 2000.

[3] L. Wang and Y. Cao, "Improved SLM for PAPR Reduction in OFDM Systems", International Workshop on Intelligent Systems and Applications, pp. 1-4, May 2009.

[4] L. Wang and J. Liu, "PAPR Reduction of OFDM Signals by PTS With Grouping and Recursive Phase Weighting Methods", IEEE Transactions on Broadcasting, vol. 57, no. 2, pp. 299-306, June 2011.

[5] Y.Z. Jiao, X.J. Liu, and X.A. Wang, "A Novel Tone Reservation Scheme with Fast Convergence for PAPR Reduction in OFDM Systems", Consumer Communications and Networking Conference, pp. 398-402, January 2008.

[6] S. Yoo, S. Yoon, S.Y. Kim, and L. Song, "A novel PAPR reduction scheme for OFDM systems: selective mapping of partial tones (SMOPT)", IEEE Transaction on Consumer Electronics, vol. 52, no. 1, pp. 40-43, February 2006.

[7] E. Bouquet, S. Haese, M. Drissi, C. Moullec, and K. Sayegrih, "An innovative and low complexity PAPR reduction technique for multicarrier systems," Proc. 9th European Conference on Wireless Technology, pp. 162-165, September 2006.

[8] C. L. Wang, Y. Ouyang, and H. C. Chen, "A low-complexity peak-to-average power ratio reduction technique for OFDM-based systems", Proc. 60th IEEE Vehicular Technology Conference, vol. 6, pp. 4380-4384, September 2004.

[9] J. Armstrong, "New OFDM Peak-to-Average Power Reduction Scheme", Proc. of IEEE Vehicular Technology, pp. 756-760, May 2001.

[10] M. Deumal, C. Vilella, J. L. Pijoan, and P. Bergadà, "Partially Clipping (PC) Method for the Peak-to-Average Power Ratio (PAPR) Reduction in OFDM", IEEE Int. Symposium on Personal, Indoor and Mobile Radio Communications, vol. 1, pp. 464-468, September 2004.

[11] X. Huang, J. Lu, J. Zheng, and J. Gu, "Piecewise-scales transform for the reduction of PAPR of OFDM signals", IEEE Global Telecommunications Conference, vol. 1, pp. 564-568, November 2002.

[12] T. Moazzeni, H. Selvaraj, and Y. Jiang, "A Novel Multi-Exponential Function-based Companding Technique for Uniform Signal Compression over Channels with Limited Dynamic Range", Intl. Journal of Electronics and Telecommunications, Vol. 56, No. 2, pp. 125-128, June 2010.

[13] X. Wang, T. T. Tjhung, C. S. Ng, and A. A. Kassim, "On the SER Analysis of A-Law Companded OFDM System", in Proc. Global Telecommunication Conference, vol. 2, pp. 756-760, December 2000.

[14] V. Cuteanu, A. Isar, "PAPR reduction of OFDM signals using hybrid clipping-companding scheme with sigmoid functions", International Conference on Applied Electronics, pp. 75-78, September 2011.

[15] B.S. Krongold and D. L. Jones, "PAR Reduction in OFDM via Active Constellation Extension", IEEE Transactions on Broadcasting, Vol. 49, No. 3, pp. 258-268, September 2003.

[16] M. Malkin, B. Krongold, and J. M. Cioffi, "Optimal constellation distortion for PAR reduction in OFDM systems", IEEE 19th Int. Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1-5, September 2008.

[17] A. Kliks and H. Bogucka, "Improving Effectiveness of the Active Constellation Extension Method for PAPR Reduction in Generalized Multicarrier Signals", SpringerLink, Wireless Personal Communication, vol. 61, no. 2, pp. 323-334, May 2010.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Hybrid Wavelet-Based Algorithms with Fast Reconstruction Features

Ileana-Diana Nicolae
Dept. of Computers and Information Technology
University of Craiova
Craiova, Romania
nicolae_ileana@software.ucv.ro

Petre-Marian Nicolae* and Marian-Ştefan Nicolae+
Dept. of Electr., Energetic and Aero-Spatial Engineering
University of Craiova
Craiova, Romania
* pnicolae@elth.ucv.ro , + nmarianstefan@yahoo.com

Abstract— **The paper is concerned with various algorithms used for the analysis based on the Discrete Wavelet Transform (DWT) of (non)stationary regimes involving almost sinusoidal waves and for data communication applications respectively. Different techniques are employed to perform analysis based on (de)compositions using two different trees: an unbalanced 10 level tree and a 6 level binary tree respectively. Our hybrid algorithms (for filters of length 4 and 6) are described and discussed. Their usability is demonstrated both for high and low power applications, a significant advantage being related to their fast "exact reconstruction" property. Considering the results and other important criteria (run time, memory consumptions), practical recommendations are made with respect to the selection of a proper reliable and fast DWT algorithm depending on the real applicability scenario.**

Keywords-discrete wavelet transform; hybrid algorithms; signals reconstruction; digital transmissions.

I. INTRODUCTION

Fourier Transforms (FT) are very helpful in signal processing, but they capture global features. They evaluate harmonic components of the entire signal, being obtained by dot-producting of the whole signal. Therefore local features can get lost and, if signal is not stationary (features change with time or in space) then this is not captured by FT. On the contrary, the wavelet transform can provides frequency information locally [9]. Wavelets have found beneficial applicability in various aspects of wireless communication systems design including channel modeling, transceiver design, data representation, data compression, source and channel coding, interference mitigation, signal de-noising and energy efficient networking [14].

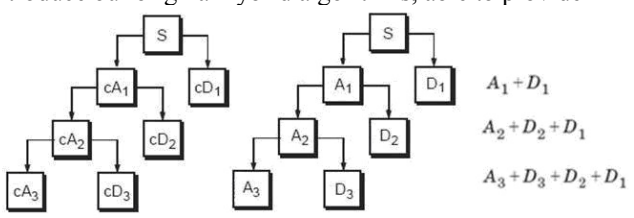
The DWT (Discrete Wavelet Transform) analyzes the signal at different frequency bands with different resolutions by decomposing the signal into an approximation containing coarse and detailed information. DWT employs two sets of functions, known as scaling and wavelet functions, which are associated with low pass and high pass filters. The decomposition of the signal into different frequency bands is simply obtained by successive high pass and low pass filtering of the time domain signal. The original signal $S[n]$ is first passed through a half-band high pass filter $g[n]$ and a half-band low pass filter $h[n]$. A half-band low pass filter removes all frequencies that are above half of the highest frequency, while a half-band high pass filter removes all frequencies that are below half of the highest frequency of the signal. The low pass filtering halves the resolution, but leaves the scale unchanged. The signal is then sub-sampled by two since half of the number of samples is redundant, according to the Nyquist's rule [14].

When the DWT is used to analyze periodic and almost sinusoidal waveforms, firstly the original waveform S is decomposed in approximations and details. Afterward successive decompositions of the approximations are made, with no further decomposition of the details (Fig. 1), based on an unbalanced tree [5]. When sets of almost random data corresponding to adjacent equally spaced frequency bands are analyzed, the wavelet packet transform is used. It is just like the wavelet transform except that it decomposes even the high frequency bands which are kept intact in the wavelet transform, using a balanced tree [12].

As Wavelets are widely applied in many scientific areas, huge efforts were done to deduce better, faster algorithms. Unfortunately, these efforts are disjointed among several disciplines [15], too few researches being interested (as this paper is) in multidisciplinary applications. Because a lot of large scientific problems require adaptivity, a collaborative framework must be developed, this involving mathematical techniques, computational methods and software [15].

A very difficult problem related to wavelet analysis is the providing of "exact reconstruction" feature, an extended study on reconstruction errors being made in [16], but no general solutions could be found.

Some of our research activities were recently dedicated to the DWT analysis of waveforms from power applications, in order to get more accurate and faster diagnosis/ evaluation methods employing specialized functions from MATLB that implement DWT [5], [6], or using our original algorithms for filters of length 4 - [8]. Our previous efforts were not concerned yet with the "exact reconstruction" property of the studied algorithms. A step forward is made with this paper, presenting our newly implemented algorithms with filters of length 6 dedicated to (non)stationary regimes. We also introduce our original hybrid algorithms, able to provide



(a) (b)

Figure 1. Signal decomposition in approximations and details (a) and its re-composition (b).

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

supplementary data, enabling our newly conceived original reconstruction functions to provide a "fast perfect reconstruction property". New research interests were revealed for us with this paper, as it contains our first studies on modern wavelet-based transmission techniques.

After a short introduction, we dedicated a section to the operational context in which we conceived and tested our algorithms and Matlab specialized functions and we decided what DWT technique is more convenient. Few mathematical fundaments are provided in Section III, to explain how the filters from our algorithms were calculated. Details on our direct and inverse algorithms with filters of length 4 or 6, using (or not) the interpolation to evaluate the missing right-end components are provided by Section IV, along with the presentation of some reconstruction-related problems. Section V is dedicated to our original hybrid algorithms, which are usable in all regimes and make possible the "exact reconstruction" property in difficult contexts. A comparative study is made relative to our algorithms' usability for the simulation and implementing of an orthogonal frequency division multiplexing system in Section VI. In Section VII, metrics are provided, that reveal the superiority of our algorithms. Conclusions and directions for our future work are presented in the end.

II. OPERATIONAL CONTEXT

The first application was implemented on a desktop (with the processor frequency of 2.4 GHz and 2GB of RAM) running Matlab vers. 7.1 under Windows XP, personalized to run for „best performances". Our data acquisition system supplies 560 samples/period per channel. The studied signal (Fig. 2), representing a phase current has distortions and is seriously affected beginning with its 5-th period.

The power quality indices considered to perform a comparison between various calculation methods are: the "node-zero" current (I_{j0}), the "non-zero node" current (I_{jn}) and the current's RMS value (I_{ef}) respectively. For an umbalanced tree with j levels and a currently analyzed number of 2^N samples, denoting by $a^{(0)}$ the approximation vector from the ultimate decomposition level (e.g. cA_3 in Fig. 1) and by $d_l^{(n)}$ the detail vector for the level l and a non-zero node (e.g. cD_i in Fig. 1), the following expression for the current's RMS value was used [11]:

$$I_{ef} = \sqrt{\frac{1}{2^N}\sum_{k=0}^{2^{N-j+1}-1}\left(a^{(0)}(k)\right)^2 + \frac{1}{2^N}\sum_{l=1}^{j}\sum_{k=0}^{2^{N-l+1}-1}\left(d_l^{(n)}(k)\right)^2} = \quad (1)$$
$$= \sqrt{I_{j0}^{\,2} + I_{jn}^{\,2}}$$

I_{j0} denotes the RMS value for the band with the lowest frequency j_0. I_{jn} represent the sets of RMS values for higher frequency bands.

The results depicted by Fig. 3 were obtained with the following methods: our original functions dwm4 and dw4i [8] implementing a DWT algorithm where different assumptions are made relative to the right border of the signal with finite length (as described in Section IV), dwt (from Matlab), considering all the values for the option "dwtmode"

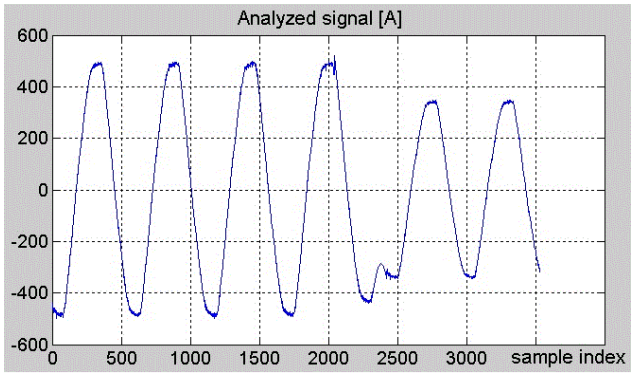


Figure 2. The analyzed signal.

that provide distinct treatments of the "finite signal's boundaries problem" [7], called with the same 4-length filter as that used by dwm4 and dw4i respectively (see Section III). The calculated filter's coefficients were found as being identical to those used by default by Matlab when dwt is called with the parameter "db2".

The analysis made with the Fast Fourier Transform (FFT) of the considered signal revealed differences lower then 1% for the power quality indices calculated with dwm4. It is why a dotted horisontal line was placed in Fig. 3 to represent the value calculated with dwm4 at the figures for stationary regime. For similar reasons dots were placed for the nonstationary regime following the value yielded by dw4i.

We applied a "shortening" procedure to the decomposition vectors obtained with dwt, when "dwtmode" was set to any of the values from the set: {"sym", "symw", "asym", "asymw", "zpd", "spd", "sp0" , "ppd", "per"}, as the internal algorithm of dwt artificially adds new components to vectors resulted from the DWT decomposition trying to rich the "exact reconstruction" and "global power preservation" properties.

We calculated the relative differences between the power quality indices with:

$$\begin{aligned} diff\_s &= (val\_s - val\_dwm4\_s)/val\_dwm4\_s\, x100; \\ diff\_n &= (val\_n - val\_dw4i\_n)/val\_dw4i\_n\, x100 \end{aligned} \quad (2)$$

where *"n"* or "*s"* denotes "nonstationary" or "stationary".

Their analysis revealed that in stationary regime, good results are provided by dwm4 and dwt with the following options for dwtmode: "ppd", "per", "asymw" and "spd". On the other hand, in nonstationary regime, dw4i is a good option while dwt can be called assigning to dwtmode the value:
- "asym" if errors of over 20% are accepted for I_{jn};
- "asymw" if errors over 10% are accepted ;
- "zpd" , as in the case of "asym", but it is not recommended in applications where more consecutive segments are analyzed to assume that the currently analyzed segment is zero beyond its borders.

The explanation for the high differences between the calculated power quality indices consists in the way of treating the "edge effect", as depicted by Figures 4 and 5. The arrows are used to mark the incorrect handling of the borders, that not only affects the calculated indices, but results also into the detection of "fake faults".

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

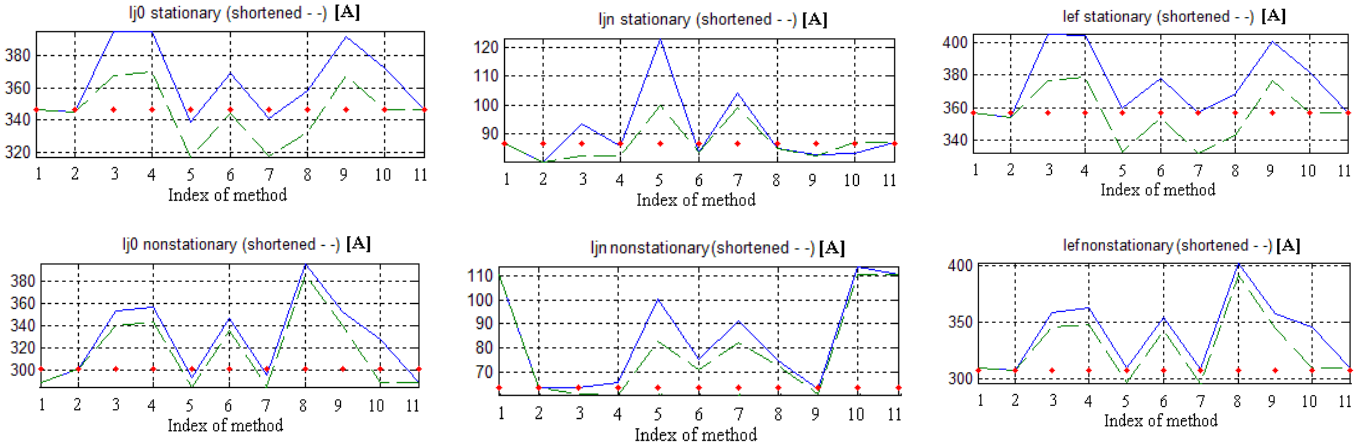


Figure 3. Power quality indices calculated with various methods. Indices for methods: 1 – dwm4; 2 – dw4i; 3…11 , dwt called with different values of dwtmode: 3- 'sym'; 4 -'symw'; 5 - 'asym'; 6 – 'asymw'; 7 – 'zpd'; 8 – 'spd'; 9 – 'sp0'; 10- 'ppd'; 11 – 'per'.

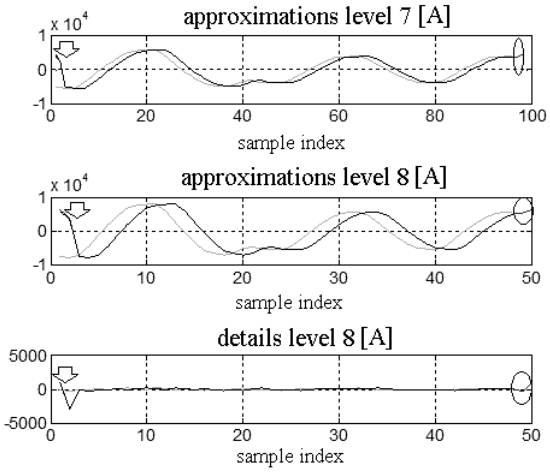


Figure 4. Nonstationary, dwt called with "ppd".

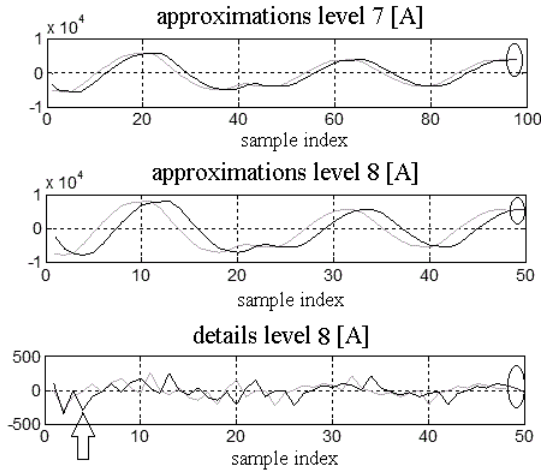


Figure 5. Nonstationary, dwt called with "asymw".

The darker waveforms correspond to the use of dwt whilst the lighter ones correspond to dw4i. The ellipses surround the components introduced artificially by dwt.

III. MATHEMATICAL FUNDAMENTS

A major problem in the development of wavelets was the search for scaling functions that are compactly supported, orthogonal, and continuous. Efforts were made to find the low-pass filter h, or equivalently, the Fourier series $H(\omega)=\sum_{k=0}^{N} h_k e^{-ik\omega}$. This trig polynomial $H(\omega)$ is often called the symbol of scaling function ϕ, which has support on $[0,N]$ [1], [4]. To ensure orthogonality, H must satisfy

$$|H(\omega)|^2 + |H(\omega+\pi)|^2 = 1,\ \omega \in \Re. \qquad (3)$$

This is true for both DWT and the continuous settings. To create a good lowpass filter, $H(\omega)$ must have vanishing derivatives at $\omega=\pi$, which forces the graph of $|H(\omega)|$ to be flat near $\omega=\pi$. Starting from the statement: „if $H(\omega)$ is the Fourier series of a low pass filter, then $G(\pi)=H(\omega+\pi)$ is the Fourier series of a high pass filter" , considering that the orthogonality constraints of G are the same as those imposed to H, and supplementary imposing that $H'(\pi)=0$ (condition introduced by Daubechy to flatten more H), in [1] and [3], the values for the filters h and g are calculated as :

$$h_0 = \frac{1+\sqrt{3}}{4\sqrt{2}};\ h_1 = \frac{3+\sqrt{3}}{4\sqrt{2}};\ h_2 = \frac{3-\sqrt{3}}{4\sqrt{2}};\ h_3 = \frac{1-\sqrt{3}}{4\sqrt{2}}; \qquad (4)$$

$$g_0 = h_3;\ g_1 = -h_2;\ g_2 = h_1;\ g_3 = -h_0. \qquad (5)$$

The filters used to reconstruct the signal s from the approximation/detail vectors obtained with the filters h and g can be expressed as [10]:

$$lh_0 = h_2;\ lh_1 = g_2;\ lh_2 = h_0;\ lh_3 = g_0; \qquad (6)$$

$$lg_0 = h_3;\ lg_1 = g_3;\ lg_2 = h_1;\ lg_3 = g_1. \qquad (7)$$

For a filter of length 6, the orthogonality conditions are [3]:

$$\sum_{k=0}^{5} h^2{}_k = 1; h_0h_2 + h_1h_3 + h_2h_4 + h_3h_5 = 0; h_0h_4 + h_1h_5 = 0 \cdot \quad (8)$$

The low-pass conditions are:

$$h_0 + h_1 + h_2 + h_3 + h_4 + h_5 = 1; h_0 - h_1 + h_2 - h_3 + h_4 - h_5 = 0. \quad (9)$$

Two derivative-related conditions are imposed now:

$$H'(\pi) = 0;\ H''(\pi) = 0 \cdot \quad (10)$$

Considering the above, the numerical values of the components from h were calculated as being [1]: h_0=0.3327; h_1=0.8069; h_2=0.460; h_3= - 0.135 ; h_4= - 0.085; h_5=0.035.

For the algorithm used in this paper, the high-pass filter g is selected as: $g = (h_5, -h_4, h_3, -h_2, h_1, -h_0)$, another possible solution being $(h_5, h_4, h_3, h_2, h_1, h_0)$ [3].

For the signal reconstruction, one can use:

$$lh_0 = h_4;\ lh_1 = g_4;\ lh_2 = h_2;\ lh_3 = g_2;\ lh_4 = h_0;\ lh_5 = g_0;$$
$$lg_0 = h_5;\ lg_1 = g_5;\ lg_2 = h_3;\ lg_3 = g_3;\ lg_4 = h_1;\ lg_5 = g_1. \quad (11)$$

In general, for a filter of even length $h=(h_0, \ldots h_L)$, the general system (GS) to be solved relies on $(L+1)/2$ orthogonality conditions, the conditions $H(0)=\sqrt{2}$, $H(\pi)=0$, and $(L-1)/2$ derivative conditions $H^{(m)}(\pi)=0$ [3].

IV. ALGORITHMS

A. Distinct algorithms for stationary and nonstationary waves

The algorithm used for the determination of approximation/detail vectors when performing a DWT decomposition (filter of length 4) of a signal s relies on [10]:

$$a_i = s(i) * h_0 + s(i+1) * h_1 + s(i+2) * h_2 + s(i+3) * h_3;$$
$$d_i = s(i) * g_0 + s(i+1) * g_1 + s(i+2) * g_2 + s(i+3) * g_3 \quad (12)$$

where the vector s contains the signal's discrete values, a and d denote the approximation/detail vectors, obtained when s is decomposed using the low-pass filter $h=(h_0\ h_1\ h_2\ h_3)$ and the high-pass filter $g=(g_0\ g_1\ g_2\ g_3)$ respectively.

When a filter of length 6 is used, two additional terms (corresponding to the last two components of the filters) appear at both a_i and d_i respectively:

$$a_i = \ldots + s(i+4) * h_4 + s(i+5) * h_5;$$
$$d_i = \ldots + s(i+4) * g_4 + s(i+5) * g_5. \quad (13)$$

For stationary waves, one can consider that the missing values beyond the right edge of the currently analysed data segment are equal to those from the left edge. So:
- for dwm4: $s(n+1) = s(1)$ and $s(n+2)= s(2)$;
- for dwm6 (our function with noninterpolated vectors and filter of length 6): $s(n+i)=s(i)$, with $i=1\ldots4$.

For nonstationary waves, our functions implement spline interpolation to evaluate the needed "$s(n+i)$" components using the rightmost components of s, as follows:
- for dw4i:
yi=spline([*n*-2 *n*-1 *n*], [*s*(*n*-2) *s*(*n*-1) *s*(*n*)],[*n*-2 *n*-1 *n n*+1 *n*+2]) . *s*(*n*+1) will be evaluated as *yi*(4) and *s*(*n*+2) will be evaluated as *yi*(5);
- for dw6i (our function with interpolated vectors and filter of length 6):
xi=spline([*n*-2 *n*-1 *n*], [*s*(*n*-2) *s*(*n*-1) *s*(*n*)],[*n*-2 *n*-1 *n n*+1 *n*+2]);
yi=spline([*n n*+1 *n*+2], [*s*(*n*) *xi*(4) *xi*(5)],[*n n*+1 *n*+2 *n*+3 *n*+4]).
s(*n*+*i*) will be evaluated as *yi*(1+*i*) for *i*=1…4.

B. Reconstruction algorithms

Using our functions id4m (or id6m), the reconstruction of the original signal s when its decomposition was done with dw4m (or dw6m) generates "almost zero" reconstruction errors, as no interpolations were involved (Fig. 6 and 7).

In the reconstruction algorithm corresponding to id4m, *temp* denotes the reconstructed signal (which is either the approximation vector from the level k, or the original signal, if $k=1$), a and d are the approximation/detail vectors from level $k+1$ and *lh*/*lg* are the low/high pass reconstruction filters. Starting from $j=3$ (the first 2 components have to be evaluated differently, because they use components with indices below the vectors' left boundaries) sets as instructions as those reproduced below are executed:

*temp(j)=a(i)*lh0+d(i)*lh1+a(i+1)*lh2+d(i+1)*lh3; j=j+1;*
*temp(j)=a(i)*lg0+d(i)*lg1+a(i+1)*lg2+d(i+1)*lg3.*

The left-most 2 components are calculated considering that *a(0)=a(length(a))* and *d(0)=d(length(d))*.

A similar algorithm is used by *idm6*, but two additional terms appear in each of the instructions (for the first line: *a(i+2)*lh4 + d(i+2)*lh5* and for the second line *a(i+2)*lg4 + d(i+2)*lg5*) .

Firstly, we conceived two functions in order to reconstruct the original signal from its corresponding approximation/detail vectors obtained with dw4i or dw6i. The missing components to the left were evaluated through spline interpolations. Unacceptable errors were obtained near the signal's left border (errors were accumulated during 2 interpolation steps - one to the right followed by another to the left). Additional operations for errors' removing should make the algorithms unusable for time-critical applications.

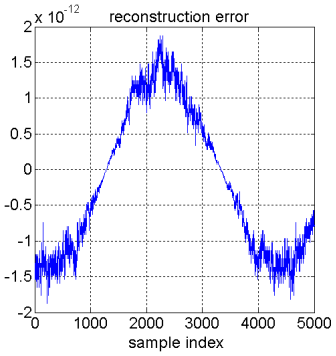


Figure 6. Stationary. Reconstruction error generated by id4m.

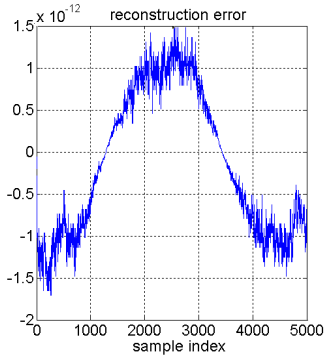


Figure 7. Nonstationary. Reconstruction error generated by id4m.

Then we applied id4m (or id6m) to reconstruct the signal from its corresponding approximation/detail vectors obtained with dw4i or dw6i. The errors were diminished (as the second interpolation was no longer performed), but still were not driven up to an acceptable level. On the other hand, when idwt (function provided by Matlab) was used to reconstruct the signal previously decomposed with dwt (providing the same value for the option dwtmode both to dwt and to idwt respectively), acceptable reconstruction errors (at most 0.1 in absolute value) were obtained for all regimes and values for the option dwtmode (Fig. 8). An exception was noticed when using the option dwtmode="per" (Fig. 9).

V. A HYBRID ALGORITHM

The reconstruction errors generated by our "idm like" functions (when applied on interpolated vectors) have something in common: they are unacceptable high for the first components of the re-constructed signal and are almost 0 afterward (order of magnitude 10^{-12}). This behavior is related to the indices of the components affected by interpolations. From Figs. 10 and 11 one can see that, except for their last components, the approximation/detail vectors (denoted by *A* and *D*) are calculated using only components from the decomposed signal (*Y*), which were never submitted to any interpolating process. For example, when d4h is used, *A*(1) and *D*(1) are calculated using only *Y*(*i*), *i*=1…4. The components affected by interpolation (either beginning with the current decomposition level or because they are calculated based on components affected by interpolations made at previous decomposition levels) are marked with *.

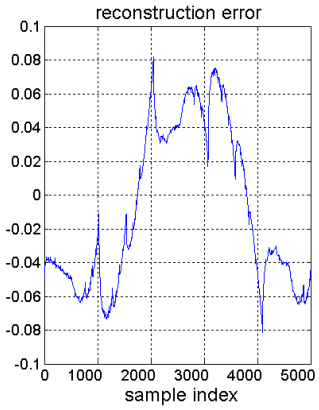


Figure 8. Filter of length 4. Reconstruction error, nonstationary regime, dwtmode="asymw".

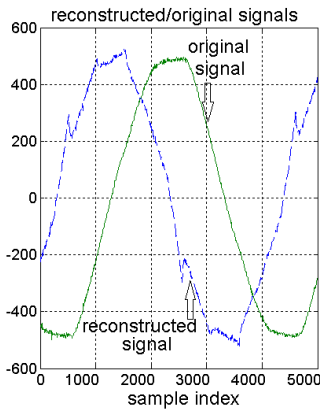


Figure 9. Filter of length 4. Reconstructed and original signals, stationary regime, dwtmode="per".

Y (1)	Y (2)	Y (3)	Y (4)	Y (5)	Y (6)	…	Y (n-3)	Y (n-2)	Y (n-1)	Y (n)	Y (1)	Y (2)
									Y* (n-1)	Y* (n)	Y* (1)	Y* (2)
A(1), D(1)												
		A(2), D(2)										
						…						
							A(n/2-1), D(n/2-1)					
							A*(n/2-1), D*(n/2-1)					
									A(n/2), D(n/2)			
									A*(n/2), D*(n/2)			

Figure 10. Conceptual schema for d4h.

…	Y (n-7)	Y (n-6)	Y (n-5)	Y (n-4)	Y (n-3)	Y (n-2)	Y (n-1)	Y (n)	Y (1)	Y (2)	Y (3)	Y (4)
					Y* (n-3)	Y* (n-2)	Y* (n-1)	Y* (n)	Y* (1)	Y* (2)	Y* (3)	Y* (4)
	A(n/2-3),D(n/2-3)											
	A*(n/2-3), D*(n/2-3)											
			A(n/2-2), D(n/2-2)									
			A*(n/2-2), D*(n/2-2)									
					A(n/2-1), D(n/2-1)							
					A*(n/2-1), D*(n/2-1)							
							A(n/2), D(n/2)					
							A*(n/2), D*(n/2)					

Figure 11. Conceptual schema for d6h.

Therefore, we conceived and tested two hybrid algorithms: d4h and d6h. d4h will be called with an instruction like: *[Cm a d]=d4h(y,n,v)*. The input parameters are: the noninterpolated version of the decomposed signal *y*, its length *n* and a vector *v* containing the last 2 components from the interpolated version of *y* (except for the first level of decomposition). The result of the decomposition performed with d4h is the vector *Cm*, whose first half represents the approximation vector and second half represents the detail vector (noninterpolated versions). The components of approximation and details affected by interpolation in the current level are provided as the vectors *a* and *d*, both having 2 components.

When a d4h-based decomposition is done for all the 10 levels, the following structures are calculated:

- 10 approximation vectors and 10 detail vectors (the noninterpolated versions). For example *cAm_k* is the approximation vector and *cDm_k* is the detail vector for the level *k* ;
- the matrix containing the final components of the approximation vectors affected by interpolation *a(2 x 10)*. For example *a(1,7)* and *a(2,7)* can be used to get the interpolated version of *cAm_7*, using two simple instructions:

cAm_7(n-1)=a(1,7); cAm_7(n)=a(2,7);

- the matrix containing the final components of the detail vectors affected by interpolation *d(2,10)*.

In a similar manner, d6h will be called with an instruction like: *[Cm a d]=d6h(y,n,v)*, but *v, a* and *d* will have 4 components instead of 2.

For the reconstruction, the noninterpolated versions of the approximation/detail vectors are used as input data to the functions that provide the signal reconstruction (id4m and id6m), according to a sequence of calls as follows:

```
cAmr_9=id4m(cAm_10,cDm_10, 2*length(cAm_10));
cAmr_8=id4m(cAmr_9,cDm_9, 2*length(cAm_9));
…
cAmr_1= id4m(cAmr_2,cDm_2, 2*length(cAm_2));
y_reconstructed=id4m(CAmr_1,cDm_1).
```

VI. APPLYING THE ALGORITHMS FOR RANDOM SETS OF DATA IN COMMUNICATIONS

Conventional OFDM (Orthogonal Frequency Division Multiplexing) systems use (I)FFT to multiplex the signals and transmit them simultaneously over a number of subcarriers. These systems employ guard intervals or cyclic prefixes (CP) so that the delay spread of the channel becomes longer than the channel impulse response. CP reduces the power efficiency, data throughput and the spectral containment of the channels.

An alternative method is known as „DWT-OFDM", that uses the Discrete Wavelet Transform to replace the IFFT and FFT blocks, and provides a better spectral containment of the channels as CP is no longer provided [12]. A typical block diagram of an OFDM system is shown in Fig. 12, where the inverse and forward transform blocks can be FFT-based or DWT-based OFDM [13].

For a DWT-OFDM system, the binary data *d* is firstly processed by a constellation mapping. A common solution for the mapping of *d* into OFDM symbols is the 16 QAM digital modulator (DM), which yields OFDM complex symbols X_m.

The analyzed scenario considers 64 channels of binary data, processed by DM in sets of 996 bits that are mapped into 166 complex symbols per set. We used the Matlab function „qammod" to obtain the symbols, that can have integer real and imaginary parts, belonging to the set {-7, -5, -3, -1, 1, 3, 5, 7}.

To obtain the time representations of both signals to be assembled in order to get the signal *y* (one for the real and the other for the imaginary parts), the 64 vectors (one per channel) formed from the corresponding symbols' real/imaginary parts were used as bottom level nodes of a binary tree. Fig. 13 depicts a 3-level tree. The number of levels for our tree was calculated as $\log_2(64)=6$. The tree can be used to apply either the inverse DWT transform („bottom-up") or the forward DWT transform („up-down"), because each channel provides data with a distinct frequency range and the band width is entirely covered by the channels' adjacent distinct frequency ranges.

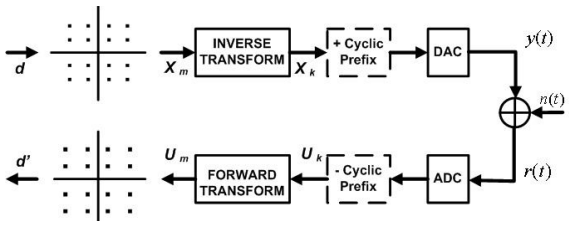


Figure 12. Schematic of an OFDM transceiver

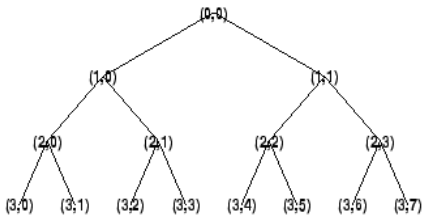


Figure 13. Example of a tree used to apply the inverse and forward DWT transforms

We firstly employed our function id4m. For example if we denote by v_{ij} the data from a node, then *v21=id4m(v32,v33,length(v32))*.

Fig. 14 depicts the first 50 discrete time values (from a total of 10624) of the signal y_r obtained from the real components , the processed data relying on 64x996 randomly generated binary data.

On the receiver side, the received data are submitted to a forward transform. The same tree can be used, with instructions like: *[v32 v33]=d4m(v21, length(v21))*.

Finally, the vectors obtained after the applying of the forward transform on the first level („received data") were compared to the data that were submitted to the inverse transform („transmitted data"). Our algorithms provided the „exact reconstruction" property. Maximum absolute differences of 10^{-13} were obtained between the transmitted and received data for all channels.

Following the same technique, but using the Matlab functions idwt, respectively dwt with the same filters and all values for the option dwtmode, considerable differences were revealed between the transmitted and received data (see the maximal differences in Fig. 15 and arrows in Fig. 16). Similar results were obtained for the imaginary components.

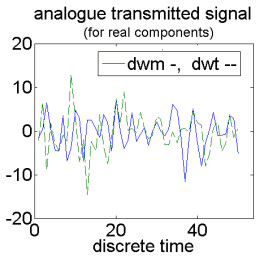


Fig. 14. The first values of the signal obtained from real components

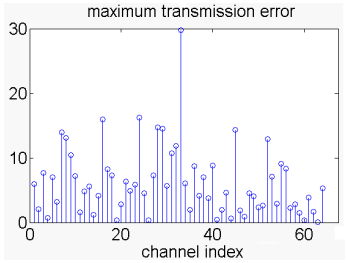


Fig. 15. Maximum transmission errors per channel (real components) when idwt/dwt were used (dwtmode='asymw')

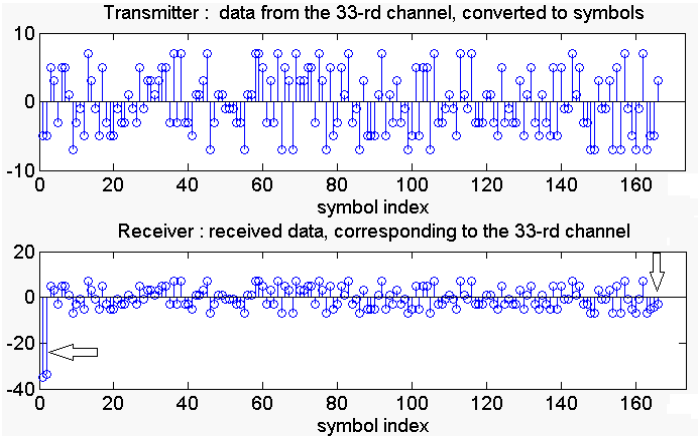


Fig. 16. Symbols corresponding to transmitted data ,real components, from the 33-rd channel (up) and received data (down), when idwt/dwt were used (dwtmode='asymw')

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

VII. METRICS

When an n – sized vector is analyzed with our hybrid algorithms using an unbalanced tree with l levels, the number of memory locations (*NML*) required to store the vectors yielded by decomposition and to restore the original signal can be calculated as the sum of :
- length of the approximation vector from the last level = $n/2^l$;
- sum of all details vectors' lengths $\sum_{1}^{l} n/2^i$;
- *2x2x l* (for d4h) or *2x4x l* (for d6h) = *NML* from the arrays used to store the values resulted through interpolations required for the estimated values from the right edge.

For a d4h-based decomposition in l=l0 levels, $NML=n+40$. For 3 periods (Fig. 2), n=12288, but n is usually larger, as usually more than 16 periods are analyzed. Therefore the additional memory requirements involved by the storing of the decomposition vectors required for the analyzed vector reconstruction are negligible as compared to the case when the analyzed vector is stored. The storing of decomposition vectors is preferred instead the original signal's storing, as the procedure to "exactly" reconstruct the original signal requires a runtime that is significantly smaller (25% from that required to determine the decomposition vectors with hybrid algorithms). Table I presents mean decomposition runtimes in this scenario.

At stationary regimes, dwm performs almost 2 times faster than dwt used with "asymw", with less memory consumption. idwt operates correctly only over "un-shortened" vectors, this making it unusable for the evaluation of power quality indices. At nonstationary regimes, dwi is slower than dwt used with "asymw", but provides accurate power quality indices and correct fault detection, with no "boundary effects". When no apriori information is available relative to the (non) stationary nature, it is indicated to use dwh as it provides vectors for all regimes (stationary or not), along with the data required to "exactly" reconstruct the signal.

In the communication application, (i)dwm and (i)dwt exhibited comparable runtimes and memory consumptions, but d4m and id4m are better options, as only they provided "exact reconstructions".

VIII. CONCLUSION AND FUTURE WORK

The Matlab toolkit for DWT analysis is not always the best option for a proper analysis of waveforms from power systems. It might introduce artificial energies of decomposition vectors, unacceptably revealed in the values of indices used for power quality analysis, supplementary memory consumption, detection of fake faults, supplementary run-times, conditioned recovery procedures (providing very poor results for the option "dwtmode"= "per"). Our algorithms do not generate longer decomposition vectors, detect correctly the moment when a fault occurs, can calculate the RMS value in nonstationary regimes for which FFT is hard to apply in real-time restrictions and exhibited good run-times in the analyzed scenarios. They can be used both in stationary and nonstationary regimes, providing a fast "exact reconstruction" method.

In communication applications, when data submitted to analysis have a random nature and can take values from a restricted set of integer values with a relative modest variation, the algorithms d4m and id4m are very good options, providing "exact reconstructions" and good runtimes, whilst the Matlab functions dwt/idwt exhibited unacceptable transmission errors.

Our future work will focus on the exploring the abilities of our algorithms in more applicability domains and their adaptation to artificial intelligence techniques.

TABLE I. MEAN RUNIMES FOR 1500 DECOMPOSITION OF 3 PERIODS

Filter length	Mean runtimes [sec]					
	asymw	dwm	dwi	dwh	idwt	reconstruction (using data provided by hybrid algorithms)
4	0.0024	0.0011	0.0031	0.0056	0.0030	0.0014
6	0.0028	0.0015	0.0054	0.0078	0.0036	0.0019

ACKNOWLEDGMENT

The work was supported by the Romanian Programe PN II "Partnership in Priority Domains", the grant "SECENGES".

REFERENCES

[1] P. Van Fleet, Discrete Wavelet Transformations: An Elemententary Approach with Applications, Wiley-Interscience, 2009.
[2] D. Percival and A. Walden, Wavelet Methods for Time Series Analysis, Cambridge , Cambridge University Press, 2006 .
[3] V.F. Patrick, Daubechies Filters, PREP 2006, Wavelet Workshop, 2006, available on line at http://cam.mathlab.stthomas.edu/ wavelets/pdffiles/UST06/Lecture6.pdf <retrieved: March, 2012>.
[4] C. Bénéteau, Haddad, C., D. Ruch and P. Van Fleet, Classical Theory and Daubechies Waveletes, PREP 2008, Wavelet Workshop, 2008, available on line at http://cam.mathlab.stthomas.edu / wavelets/pdffiles/UST08/Lecture7.pdf, <retrieved: March, 2012>.
[5] I.D. Nicolae and P.M. Nicolae, Using Wavelet transform for the evaluation of power quality in distorting regimes, Acta Electrotechnica, vol. 52, no .5, pp. 331-338, 2011.
[6] I.D. Nicolae and P.M. Nicolae, Using discrete Wavelet transform to evaluate power quality at highly distorted three-phase systems, Proceed. of the 11-th Int. Conf. on Electrical Power Quality and Utilization (EPQU 11), pp: 1 – 6, 17-19 Oct. 2011, Lisboa, Portugal, doi: 10.1109/ EPQU.2011. 6128825.
[7] http://www.mathworks.com /help/toolbox/wavelet/ug/f8-25097.html, <retrieved: March, 2012>.
[8] I.D. Nicolae and P.M. Nicolae, Real-time analysis using Discrete Wavelet Transform in power systems, EPE-PEMC 2012, in press.
[9] R. Ashis, Transforms, Fourier and Wavelets, available at http://www.cs.cornell.edu/ courses/cs5540/2010sp/lectures/Lec5.Transforms.pdf , <retrieved: March, 2012>.
[10] I. Kaplan, The Daubechies D4 Wavelet Transform, available at http://www.bearcave.com/software/java/wavelets/daubechies/index.htm, <retrieved: March, 2012>.
[11] Morsi, W.G. and El-Hawary, M.E., Wavelet Packet Transform-Based Power Quality Indices for Balanced ad Unbalanced Three-Phase Syst. under Stationary and Nonstationary Operating Conditions, IEEE Trans. on Power Delivery, vol. 24, no. 4, pp. 2300-2310, 2009.
[12] R. Dilmirghani and M. Ghavami, „Wavelet Vs Fourier Based UWB Systems", 18t-h IEEE Intern. Symposium on Personal, Indoor and Mobile Radio Communications, pp.1-5, 2007.
[13] S. Baig, F.Farrukh and M. J. Mughal, Discrete Wavelet Transforms - Algorithms and Applications, Intech, 2011.
[14] M. K. Lakshmanan and H. Nikookar, A Review of Wavelets for Digital Wireless Comm., Wireless Personal Communications, vol. 37, No. 3-4, 387-420, DOI: 10.1007/s11277-006-9077-y.
[15] G. Latu, Data Structure Design and Algorithms for Wavelet-Based Applications, available at http://icps.u-strasbg.fr/people/latu/ public_html/wavelet/course_slide.pdf, 2010, <retrieved: March, 2012>.
[16] T. K. Sarkar, M. Salazar-Palma and M.C. Wicks, Wavelet Applications in Engineering Electromagnetics, Artech House, 2002.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

DMT: A new Approach of DiffServ QoS Methodology

Rashid Hassani, Amirreza Fazely
Department of Computer Science
University of Rostock
Rostock, Germany
rashid.hassani@uni-rostock.de
amirreza.fazelyhamedani@uni-rostock.de

Peter Luksch, Abbas Malekpour
Department of Computer Science
University of Rostock
Rostock, Germany
peter.luksch@uni-rostock.de
abbas.malekpour@uni-rostock.de

***Abstract*—Quality of service (QoS) refers to the ability to provide guarantees w.r.t. to bandwidth, latency, jitter, etc., to certain classes of network traffic. The effectiveness of QoS strategies and their implementation depend on a large number of factors, e.g., the size of the network and the complexity of services the network is intended to provide to users. In this paper, we propose a layered QoS which guarantees that the available bandwidth is assigned to users proportionate to the subscribed bandwidth even in case of congested backbone links. The key issue to achieve this is effective prioritization of management traffic. We have implemented our QoS strategy in a laboratory environment and have monitored its performance under simulated traffic. Our method has significantly reduced the total amount of packet loss. Bandwidth utilization on the congested link was increased by 60 percent.**

Keywords-QoS; traffic management; DiffServ; DMT; BGP.

I. INTRODUCTION

By considering the future of the Internet, it will be seriously overburdened by different traffic sources such as real time traffic (e.g., video/voice) and huge traffic generated by e-commerce transactions. In this variety of traffic, network congestion is a concern that can bring different problems to any data network. This issue is even more serious when data network should be accessed, managed and monitored from distance. Mostly, on each ISP network, the essential traffic may be considered as management traffic for remote access and controlling the network devices and traffic of voice/video. In order to design a network properly to support management traffic, QoS mechanisms require to be implemented in order to guarantee that management traffic is prioritized properly.

QoS is a technique to prioritize certain classes of traffic while at the same time maximizing resource utilization. It cannot increase the bandwidth capacities, but by using QoS, network administrator priorities the traffic in a way that if a link is congested, they could choose purposely to drop lower priority traffic so the higher priority traffic will be gradually served. Therefore, QoS doesn't help to avoid dropping the traffic, but it can facilitate the traffic flow in such a way that sensitive traffic continually is serving the network.

Management traffic must have highest priority, because resource management, recovery from failure, and other management services can only be effective if they can reliably and quickly reach each device at any time. Integrated Services (IntServ) and Differentiated Services (DiffServ) [2] are two architectures that have been developed by IETF for applying QoS in IP-networks. Based on the researches, DiffServ so called flow aggregation model can offer the same or even better QoS than the reservation based model [6] (i.e., IntServ).

There are many DiffServ QoS techniques available which have been investigated by several projects e.g., RMD [10], IntServ over DiffServ [11], Bandwidth Broker [12] and Pre Congestion Notification (PCN) [13]. RMD has two ways to control network traffic. The first is to control the flows entering the network and the second one is an algorithm that terminates the required amount of flows if the network is congested within the domain. RMD was developed to provide dynamic QoS within a DiffServ network in a scalable way. IntServ over DiffServ is an end-to-end QoS which is applied by using the IntServ model across a network with one or more DiffServ areas. A Bandwidth Broker (BB) is a centralized agent which has information about the bandwidth precedence and policies in a network and assigns bandwidth by considering those policies. There are other architectures that use this technique (BB) for example: the TEQUIL [14], AQUILA [15] and Internet2 QBone [12]. The Pre Congestion Notification (PCN) is a DiffServ technique in which the PCN-enabled Interior nodes try to detect congestions.

Along with consideration of above techniques, we looked for a simple and cost-effective solution to be applicable for the ISP's backbone network especially for the region in which limited bandwidth, highly flooded real time network traffic, network device overhead and lack of redundancy are vital points for that ISP.

We propose a new QoS strategy which is based on the differentiated services mechanism (i.e., a standardized mechanism of classifying and managing network traffic). In our solution, the traffic that enters a DiffServ domain from outside is classified by marks. It has to be decided to what extend this classification is *trusted*. As we use _D_iffServ

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

with _M_arking and _T_rusting on the backbone, we call our method as **DMT**.

Section II consists of problem statement and its proposed solution. Section III describes the scenario and its implemented procedure in details. Finally, Section IV and V present experimental results and conclusion respectively.

II. PROBLEM STATEMENT AND PROPOSED SOLUTION

In general, this paper tries to provide a solution for the following problem: The ISP which we consider is located in the geographical area in which the ISP's internet bandwidth might be very expensive and the users may experience the lack of bandwidth. The inadequate internet links are regularly congested so the total bandwidth of an area in congested times is distributed to the users proportionate to the subscribed bandwidth. Therefore regarding management and maintenance issues, ISP suffers from lack of bandwidth and redundancy [4]. The resulting solution must fit with the hardware capabilities of the devices, which are used in the network. In this work, there were some vital points that are essential to be considered, such as:

- It should be considered where management traffic originally must be marked, or classified, and which devices should do the marking.
- Different network devices have different traffic management capabilities. Therefore to conquer these differences there should be a way to employ an ordinary packet marking policy.
- When network traffic is transferred between the LAN and WAN, it must be determined how to map marking policies between OSI Layer-2 (Data Link) and Layer-3 (Network) levels.

In order to find and apply the proper QoS mechanism on the network, we considered different solutions. The solution which could provide all requirements, prioritizes the overall traffic flows to ISP nodes in order to facilitate remote management in the network is chosen to be proposed here. Other solutions are either unsuccessful in experiments or incoherent in the existing topology of the network.

III. DMT APPROACH

Nowadays, the routing architecture specially used in internet is based on the *best effort* communication model [8]. For some kinds of traffic like web, best effort is regularly good enough but it does not guarantee actual delivery or timeliness [3]. If packets get lost somewhere on their way to their destination, the end hosts (senders) must retransmit the missing packets. However, specific packets like management traffic require better performance; therefore, this kind of traffic should be considered as the highest possible priority.

The Internet consists of thousands of various networks which are managed and controlled by either a single administrator or institution that is called Autonomous System (AS). BGP (The Border Gateway Protocol) is the routing protocol designed to exchange information between these ASs [5]. There are other routing protocols such as IS-IS, RIP and EIGRP but they cannot operate and be used in the same way and for the same purpose that BGP does [9]. Therefore for this scenario the BGP is employed as a routing protocol between different routers and switches as necessitate of ISP to manage where broadcast packets have to be forwarded.

DiffServ is a widely used networking architecture mechanism for traffic management and provides QoS on modern IP networks. *DiffServ* operates on the standard that is called traffic classification by placing each data packet into a limited number of traffic classes. DiffServ uses the 6 most significant bits field in IP packet header which is called DSCP. DSCP bits are used instead of TOS (Type Of Service) field which is now outdated.

As shown in Figure 1, DiffServ Field has 8 bits which are separated in to two parts; one DSCP with six bits (DS5-DS0) and second ECN with two bits. In the real networks with DiffServ as a QoS mechanism, each packet is marked by using the DiffServ field so that it is given at each network node a specific forwarding behavior.

DS5	DS4	DS3	DS2	DS1	DS0	ECN	ECN

Figure 1. DiffServ field

In the architecture of DiffServ, a field called DiffServ field (DS) has been defined, which is replaced the TOS field in IPv4. It is used to decide about packet classification and different traffic conditioning purposes such as metering, policing, marking and shaping.

Table I shows different Precedence Levels which are shown by DSCP decimal for each level.

TABLE I: DSCP different Precedence Levels

Precedence Level	DiffServ Marking	Description
7	DSCP 56 (CS7)	Used for link layer and routing protocol keep alive
6	DSCP 48 (CS6)	Reserved for IP routing protocols
5	DSCP 40 (CS5)	Express Forwarding (EF)
4	DSCP 32 (CS4)	Class 4
3	DSCP 24 (CS3)	Class 3
2	DSCP 16 (CS2)	Class 2
1	DSCP 8 (CS1)	Class 1
0	DSCP 0 (Default)	Best Effort

The usage of the DSCP field can be categorized in to three ways:

- Classifier: Choose a packet by considering the contents of some parts of the packet header and by using the predefined DSCP value.
- Marker: By considering the traffic profile, it will set the DSCP field value.
- Metering: By using the sharper or dropper function, it will check the fulfillment of traffic profile.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

The following scenario will clarify the above-mentioned problem.

A. Scenario

The Figure 4 represents the network topology, which is planned for this scenario. The main concern is to provide the way to overcome the management traffic problem in congested links with limited bandwidths. Therefore in our Lab scenario, the links with different bandwidths and network devices are only considered in order to simulate a real world network backbone environment. Other network issues such as routing are not taken in to the consideration.

In order to generate management traffic, we have used two computers and specific network management applications. Different capacities for the used links have been configured and some bottlenecks have been made to observe the efficiency of the QoS mechanism for the congested links. The computer on the left acts as a management server to generate various traffic. We have provided remote access to each device in the network such as routers, switches and etc., in order to monitor, control and manage the network traffic. Some of these operations such as configuration or monitoring could be done either manually (i.e., *telnet* protocol) or planned and implemented automatically by using particular network management software (e.g., Solarwinds application) [1]. Therefore, we have simulated an ISP core network to test guarantee delivery for the management traffic. In order to be sure about service consistency, different types of devices have been used in ISP core such as Cisco/Juniper routers and Switches.

B. Procedure

Before considering actual traffic, we have to classify the important traffic by marking them using DiffServ mechanism. In ordinary way of DiffServ implemented (only Marker) the DSCP and traffic precedence should be defined on all the routers and switches between source and destination. When this QoS method is applied over heavy traffic, the network may experience a very high CPU load on network devices on the path especially when a variety types of traffic have been flooded to them. Therefore, some of the devices may not be able to apply appropriate QoS mechanism on the traffic. However, in our solution, we specified a minimum bandwidth in the links for management traffic on all hops. The management traffic which is created from specific management application (Solarwinds) destined to some particular destinations is marked only at the first hop and then is ***trusted*** by all middle hops. Therefore, as one of the links became congested, the minimum bandwidth over the whole path is preserved for the management traffic.

As shown in Figure 4 the traffic generator client acts as a management server to generate various traffic in particular port such as port 23 for *telnet*. By using the *DiffServ* QoS, the traffic should be marked at the first router and the mark is proceeding along traffic to reach to the destination device. The traffic which is originated from/destined to these computers is marked as '*DSCP CS1*' (Differentiate Service Code Point, first precedence) by using DiffServ method for packet classification. The RFC 5865 "Configuration Guidelines for DiffServ Service Classes" was used as instruction to set the DSCP bits in the IP packet. DSCP defines the relative priority and drop precedence for IP packets in a network. As far as the minimum bandwidth of the link is considered for the management traffic, the management server is simply able to communicate with all devices in the network in spite of any links on the path congested or not. Without this mechanism, this communication will be tremendously slow and undergoes some problems when a link becomes congested.

To examine this scenario, a traffic generator on the management server is installed to create various familiar network traffic such as IP/ICMP/TCP/UDP to congest the links. To make a better conclusion on the QoS result after and before applying it to the traffic, the ping command is used to monitor and clarify the result. So the devices are configured to select the ICMP traffic.

At the first hop, an access-list is configured in the router (Router1-2821) to select which traffic, QoS mechanism should be applied on it. This configuration is applied only once at the first hop. On the middle hops, the other classes are configured supposing that the management traffic is generated and already tagged by *dscp CS1* elsewhere not close to the first hop. This traffic passes through these routers as middle hops so that the required QoS function should be applied on it. The class is used at the same policy-map ultimately; the policy map should be applied on the egress interface of the router as output service policy. On the switch, the configuration undergoes a slight difference. The switch is configured to ***trust*** the DSCP values on some certain interfaces connected to the core network otherwise it resets the DSCP values by default. The bandwidth capacity for the link between the Cisco router (core 7200) and the Juniper is only 2mbps and therefore can be considered as the bottleneck of the network. So when the traffic flows from client to the server (from left to right), the majority of the packet drops happen in Csico router (core 7200). The numbers of packets marked in the first hop (Router1-2821) for the management server are significantly high since the *telnet* and ICMP traffic are generated here to different destinations. When all devises on the network are configured as mentioned above, a traffic generator program which is installed on the Management server starts to generate some TCP traffic along with management traffic to make the connected links to be congested. Meanwhile, to observe the QoS result on the selected traffic, the ping result is monitored on the end-to-end computers.

IV. EXPERIMENTAL RESULTS

To prepare network devices for the test and to choose the management traffic in order to apply the QoS on it, at the router Router1-2821, an access-list is configured and commands are applied only once and at the first hop, as one can see in Figure 2.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

```
ip access-list extended MGMTRF_TEST_QOS
permit tcp host 192.168.208.252 any eq telnet
permit tcp host 100.100.100.1 eq telnet host 192.168.208.252
permit icmp any any
```

Figure 2. Defined access list

As Figure 4 shows, the IP address of the host which acts as management server is '192.168.208.252'. Therefore, in Figure 2, the second line selects and permits the traffic originated by this host and the third line selects and permits the returning traffic. Fourth line is configured to choose and permit the ICMP traffic.

As one can see in Figure 3, to apply QoS function, the following class/policy-map is defined on the selected traffic.

```
class-map match-all MARK_MGMTRF
match access-group name MGMTRF_TEST_QOS
!
policy-map MGMTRF_OUT
class MARK_MGMTRF
bandwidth 20
set dscp CS1
```

Figure 3. Marking the selected traffic

After all configurations done on the network devices and generating heavy traffic, the ping is performed in one of the computers (management server) to the router. The result of the above test could reveal how successful the solution is. We specified higher priorities to the ICMP packets in order to facilitate the test. The ping times are monitored before and after applying DMT methodology (our proposed solution) for QoS mechanism. Without DMT, result shows that the average ping time never significantly falls from high level (approximately 350ms). However, after applying the DMT methodology on the selected traffic, the ping time significantly drops from 370ms to approximately 6ms (improved by about 85%).

There are several protocols and applications which can be used to create management traffic flow like Telnet, FTP, SMTP, etc. We examined *Telnet* operation, because it performs two important functions: first, it interacts with the user terminal on the local host and second, exchanges messages with the destination Telnet host (i.e., network devices). During the *Telnet* operation, the TCP connection continues for the whole period of the login session. The client and the server retain the connection, even while the user disrupts the transfer of data [7].

Before applying the DMT methodology on the congested link, the *telnet* from the management server to the routers worked extremely slowly; but, after applying the DMT methodology, the *telnet* operations seemed very usual.

Figure 5 and Figure 6 demonstrate the tests we have done to examine the simulated network packet loss before and after applying the DMT methodology respectively. By comparing the graphs, before applying DMT methodology (Figure 5) the average packet loss was extremely high in congested links (up to 85%) but after applying DMT in the network (Figure 6), the loss rate has been significantly reduced (about 25%). Therefore, the bandwidth utilization for defined traffic is highly optimized (about 60 percent).

V. CONCLUSION AND FUTURE WORK

The congestion problem on the low bandwidth links under the heavy traffic situation cannot be avoided. Serving management and maintenance services on the backbone networks over the congested links is a main factor. The defined scenario in this paper demonstrates the effectiveness of the DMT methodology in ISP's backbone network especially for the region in which limited bandwidth, highly flooded real time network traffic, network device overhead and lack of redundancy are vital points for ISPs. Therefore, DMT can be considered as a cost-effective and profitable *network based policy method* for reliable delivery of QoS by service providers. DMT methodology can be applied on any devices in the core network routers and layer-3 switches when the DSCP field on each packet remains unchanged through the path from base to destination. However, while the DSCP field is unrecognizable on some Layer-2 switches, this QoS mechanism will fail at these devices. Other defined QoS methods for managing the management traffic such as matching the L-3 traffic against the IP access-lists or assigning a fixed bandwidth to the management traffic, either decline the performance or they would be considerably costly comparing to the DMT method.

There are many interesting avenues for future work. In this paper, we have proposed a layered model of policy-based DiffServ QoS management system. We are currently at the stage of implementation this technique in a real more complex multi-domain environment with Linux-based router and also demonstrate the system on laboratory test beds.

Additionally, wireless link capacity is typically a limited resource that requires to be used efficiently. Therefore, it is important to find efficient technique of supporting QoS over wireless channels for real-time data (e.g., live audio/video streams) when capacity of the channels vary for different users. We plan to report outcome on detailed features of the proposed implementation model in future papers.

ACKNOWLEDGMENT

The authors are very grateful to the reviewers for their useful comments and suggestions.

REFERENCES

[1] Solarwinds Technologies, www.solarwinds.com, [retrieved: May, 2012].

[2] J. Polk, and F. Baker, "Configuration Guidelines for DiffServ Service Classes", RFC 5865, 2010.

[3] N. Ye, E.S. Gel, and X. Li, "Applying scheduling rules from production planning", Computers & Operations Research, Volume 32, Issue 5, May 2005, pp. 1147–1164.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

[4] D.J. Songhurst and P.L. Eardley, "Guaranteed QoS Synthesis for admission control with shared capacity", BT Technical Report TR-CXR9-2006-001, Feb 2006.

[5] Y. Rekhter, and S. Hares, “A Border Gateway Protocol 4 (BGP-4)”, RFC 4271, 2006.

[6] J. Haju and P. Kivimaki, “Co-operation and comparison of DiffServ and IntServ: performance measurements”, ISBN:0-7695-0912-6, 25th Annual IEEE Conference on Local Computer Networks, 2000.

[7] D. Comer, “Internetworking With TCP/IP, Principles, Protocols and Architecture”, ISBN 0-13-187671-6, 5th edition, 2006.

[8] B. Smith and J.L. Aceves, “Best-Effort Quality-of-Service”, ICCCN’08, Aug 2008.

[9] M. Caesar and J. Rexford, “BGP routing policies in ISP networks”, IEEE Network, November/December 2005.

[10] Attila. B, G. Karagiannis ,and L. Westberg, “Qos signaling across heterogeneous wired/wireless networks”, QShine2005, p.p. 51.

[11] Y. Bernet, P. Ford, and R. Yavatkar, “A framework for integrated services operation over diffserv networks”, IETF, Nov 2000.

[12] B.Teitelbaum, S. Hares, and L. Dunn, “Internet2 qbone: Building a testbed for differentiated services”, IEEE Network, Oct 1999, pp. 8–16.

[13] B. Briscoe, P. Eardley, and D. Songhurst, “An edge-to-edge deployment model for pre-congestion notification: Admission control over a diffserv region”, IETF, June 2006.

[14] E. Mykoniati, C. Charalampous, and P. Georgatsos, “Admission control for providing qos in diffserv ip networks: The tequila approach”, IEEE Communications Magazine, Jan 2003, pp. 38–44.

[15] T. Engel, H. Granzer, and M. Winter, “Aquila: Adaptive resource control for qos using an ip-based layered architecture”, IEEE Communications Magazine, Jan 2003, pp. 46–53.

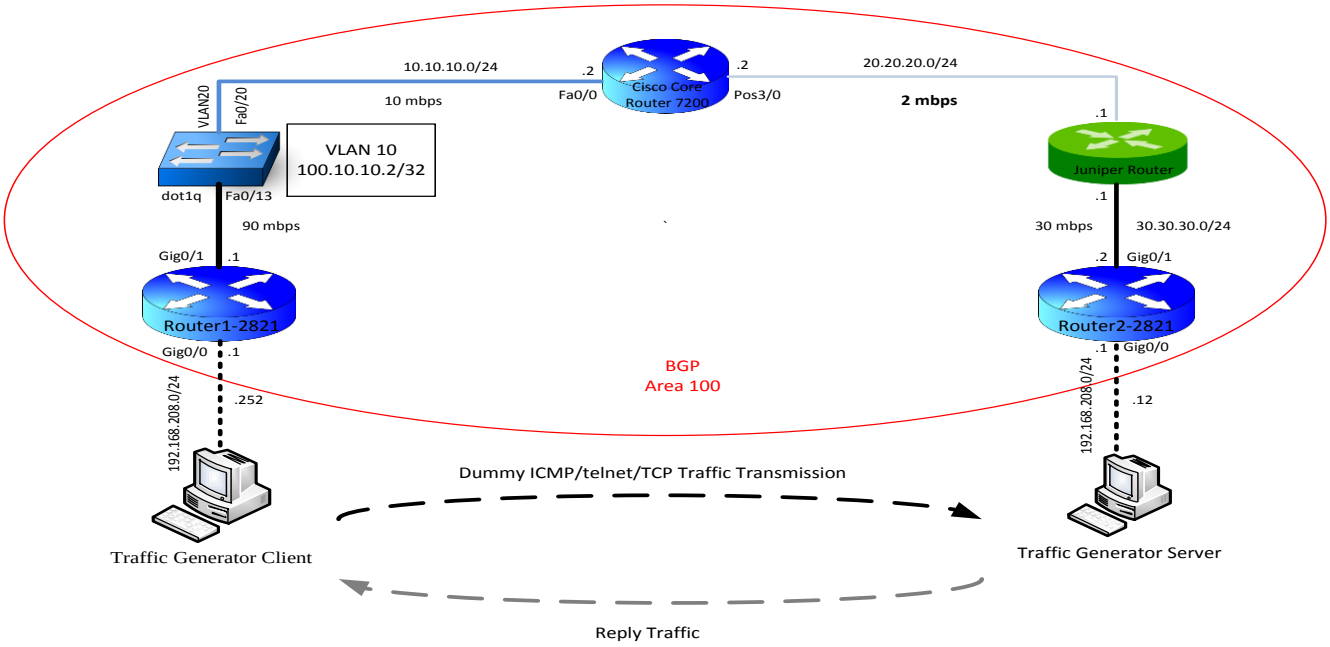


Figure 4. Sample network topology for proposed scenario

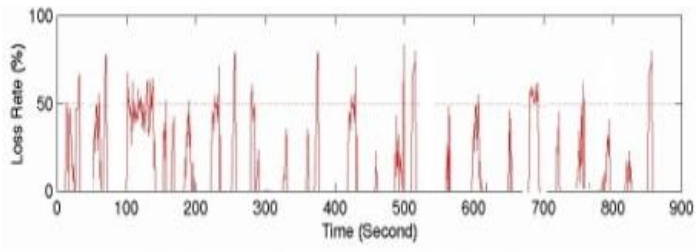


Figure 5. Packet lost before applying DMT

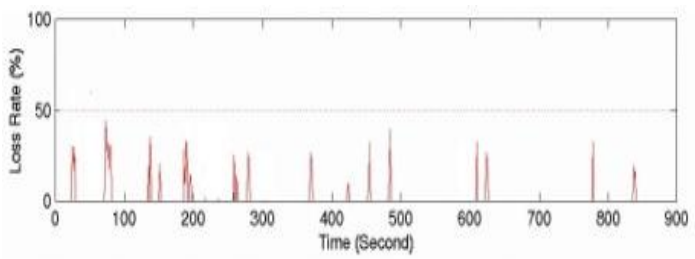


Figure 6. Packet lost after applying DMT

Fast Network-Based Brute-Force Detection

Robert Koch, Gabi Dreo Rodosek
Universität der Bundeswehr
Institut für Technische Informatik
Neubiberg, Germany
[Robert.Koch, Gabi.Dreo]@UniBw.de

***Abstract*—Today, the Internet is a crucial business factor for most companies. Different traditional business divisions like distance selling or money transfers enhanced or even switched to the Internet, others emerged directly from it and a billion dollar business evolved over the past years. Therefore, the high fiscal values are alluring criminals. Attacks with the aid of the Internet can be executed from a safe distance, different (or even missing) IT laws in different countries are hampering the transboundary criminal execution. For example, brute-force attacks to gain access to systems and servers are still a popular and successful attack type. After gaining access, sensitive data can be copied, spyware can be installed, etc. Current protection mechanisms require extensive administration or can reduce network performance. Therefore, we propose a new architecture for network-based brute-force detection in encrypted environments. The system evaluates the similarity of the network packet payload-sizes of different connections. No information about the encryption in use or the functionality of the authorization process is required. Based on the high similarity of rejected connections, an identification of brute-force attacks is realized.**

***Keywords-brute force*; *intrusion detection*; *network-based*; *similarity*; *inherent knowledge*.**

I. INTRODUCTION

In 1994, the first websites for online banking and the online ordering of pizza were published. Since then, endless services have been developed and also traditional services like distance selling moved to the Internet. The collapse of the dotcom bubble in 2000 did not end this trend, but reduced the superelevated expected profits to more realistic levels again. For example, in Germany, the World Wide Web became the revenue driver with 53.3 percent of the entire mail order business in 2010.

The financial power and the possibility to conduct attacks over the Internet from long distances (therefore, hampering a criminal prosecution even in the case of a detection and localization of an attacker) attracts numerous criminals. By that, a complex and surprisingly specialized cyber crime market has been evolved over the past years. Today, professional attack tools are dealt in the digital underground which are able to perform effective attacks.

Brute-force attacks are a popular technique to gain system access [1]. They are easy to accomplish and will have a great impact, if the attacker breaks into an account. Therefore, attacks like *SSH* brute-force are still one of the most frequent attacks [2], [3]. For example, the statistics of ATLAS provide information about the most frequent attacks in the last 24 hours [4]. Repeatetly, brute-force attacks (*SSH brute-force login attempts*) had been in the Top 5 of the most frequent attacks per subnet. Numerous tools provide an easy and automated attack conduction with detailed configuration options (e.g., number of parallel connections, number of connection tries per second, ciphering, dictionary-based or systematic tries). `SSHatter` [5] and `brutessh` [6] are two examples of corresponding attack tools.

The popularity of this kind of attack is based on the bad usage of passwords, too. Often, passwords are very simple, used for multiple systems and logins or they are created based on simple formulation rules (e.g., the current month and a number) when systems force regular changes of passwords. Different studies (e.g., see [7], [8], [9]) demonstrate that this situation does not improve even with intense information and warnings to the users.

Today, numerous services offer remote access or they are used as substructure for further services, e.g., the safeguarding of remote desktop access by tunneling unencrypted or vulnerable protocols like *RDP* over secure connections, for example by the use of *SSH*. Also, numerous web services are based on an authentication by an username-password combination which can be attacked by brute-force dictionary attacks. Therefore, the user remains a crucial weakness and opens up important points of attack. Current systems for the detection and prevention of brute-force attacks often have to be installed and administrated for every system with corresponding services. Network-based solutions often reduce network performance, too.

We present a new network-based architecture for the detection of brute-force attacks. The system evaluates statistical data of the network-connections and calculates the similarity of authorization requests to detect attacks. Neither the location or the type of services nor any information about the encryption in use is needed: The system can be placed into the network and runs out-of-the-box, resource-saving and without any configuration.

The remainder of the paper is organized as followed: In Section II, the mode of operation of remote sessions is presented and detection possibilities of brute-force attacks are identified. The architecture and the new detection principle

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Table I
PAYLOAD SIZES OF SERVER AND CLIENT PACKETS DURING AN AUTHORIZATION PROCESS. ON THE LEFT SIDE THE AUTHENTICATION FAILED, WHILE THE PACKET SERIES ON THE RIGHT WAS SUCCESSFUL: THE PROCESS FINISHED AND A SESSION WAS ESTABLISHED. n IS THE NUMBER OF THE OBSERVED PACKET, ZERO-SIZED PACKETS AND ADMINISTRATIVE NETWORK PACKETS (E.G., AN *Acknowledge*) ARE LEFT OUT FOR BETTER READABILITY.

n	Size	Source	n	Size	Source
5	39	S	5	39	S
7	39	C	7	39	C
9	792	C	9	792	C
10	784	S	10	784	S
13	24	C	13	24	C
14	152	S	15	152	S
16	144	C	17	144	C
17	720	S	18	720	S
18	16	C	19	16	C
20	48	C	21	48	C
22	48	S	23	48	S
23	64	C	24	64	C
25	64	S	26	64	S
27	144	C	28	144	C
29	64	S	30	32	S
31	144	C	32	128	C
33	64	S	34	48	S
35	144	C	35	448	C
37	64	S	37	112	S
			38	368	S
			40	80	S
			42	48	C
			43	176	S
			44	64	S
			47	32	C

is presented comprehensively in Section III, while the results of the prototypical implementation are given in Section IV. A brief overview of related work is given in Section V, while Section VI concludes the paper.

II. REMOTE SESSION BASICS

Todays remote session services are typically encrypted or transported over encrypted channels, for example *RDP* through an encrypted *SSH*-tunnel. Therefore, only limited information about the transported datagram can be evaluated and used for the detection of attacks, namely the packet sizes of the encrypted payload and the points in time when the different network packets of a session pass a specific observation point, e.g., a border gateway in the subnet. In particular, the often used content of the payload is not available. Therefore, a deep packet inspection (DPI) is not possible; a detection must be independent from the availability of the payload.

After the client initiated the connection to the server, the authentication phase is started, e.g., by the request of an username and password. After a successful authentication, the user will be able to proceed with the remote session, otherwise the login information will be requested again. Typically, the login information can be entered a limited number of times (e.g., three times), after that the server will close the session and disconnect the client. Therefore, an attacker can try to infiltrate the system by testing various username-password combinations, e.g., based on dictionaries.

The observable parameters of a *SSH* remote session are shown in Table I. In the example, the encryption algorithm `AES128-CBC` is used, therefore generating the characteristic payload sizes as shown.

As one can see, there are typical similar packet series in the case of the rejected authentication tries and different characteristic packet series in the case of a success. Therefore, these packet series can be used to provide a detection of brute-force attacks, based on the recurrent observability of packet series corresponding to rejected authentications. Anyway, because the exact sizes can differ keenly, e.g., based on the used encipher algorithm or used padding, a detection based on a comparison of the strict packet sizes would need numerous patterns in a database and one can not be sure to catch all possible cases.

III. A NEW CONCEPT FOR BRUTE-FORCE ATTACK DETECTION

To enable a network-based fast brute-force detection which is independent from the used remote session protocol as well as the used cryptographic algorithm, we present a new detection concept and corresponding detection system which overcomes the shortcomings of current systems.

The detection of attacks is realized by monitoring the encrypted datastreams and evaluating selected packet sequences by the calculation of their similarity: If the packet sequences of failed login attempts are examined by a similarity measure, e.g., a cross-correlation, repeated negative authentications will result in high similarity values. On the other side, if a connection is authenticated successfully, the packets sent for the authentication process and the subsequent network packets will be quite different, therefore ending in small similarity values. By that, the similarity between data streams of one or different connections can provide information about an ongoing attack. Two basic types for the comparison of data streams and calculating their similarity can occur, by ports (different connections of one or more IP addresses) and by lines (one or more authentication tries within one connection). Analyzing a remote session in more detail, multiple cases of parallel and serial data streams can occure (see Figure 1) and must be taken into consideration for the calculation, namely:

- Initiation of a connection and successful authentication: After the preamble, a successful authentication is fulfilled and a data connection is open up and used for data transfer.
- Initiation of a connection and repeated rejected authentication attempts. Termination of the connection after a maximum number of false tries, initiation of a new connection on a new port.
- Initiation of parallel connections with one or more authentication attempt(s), termination by the client after

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

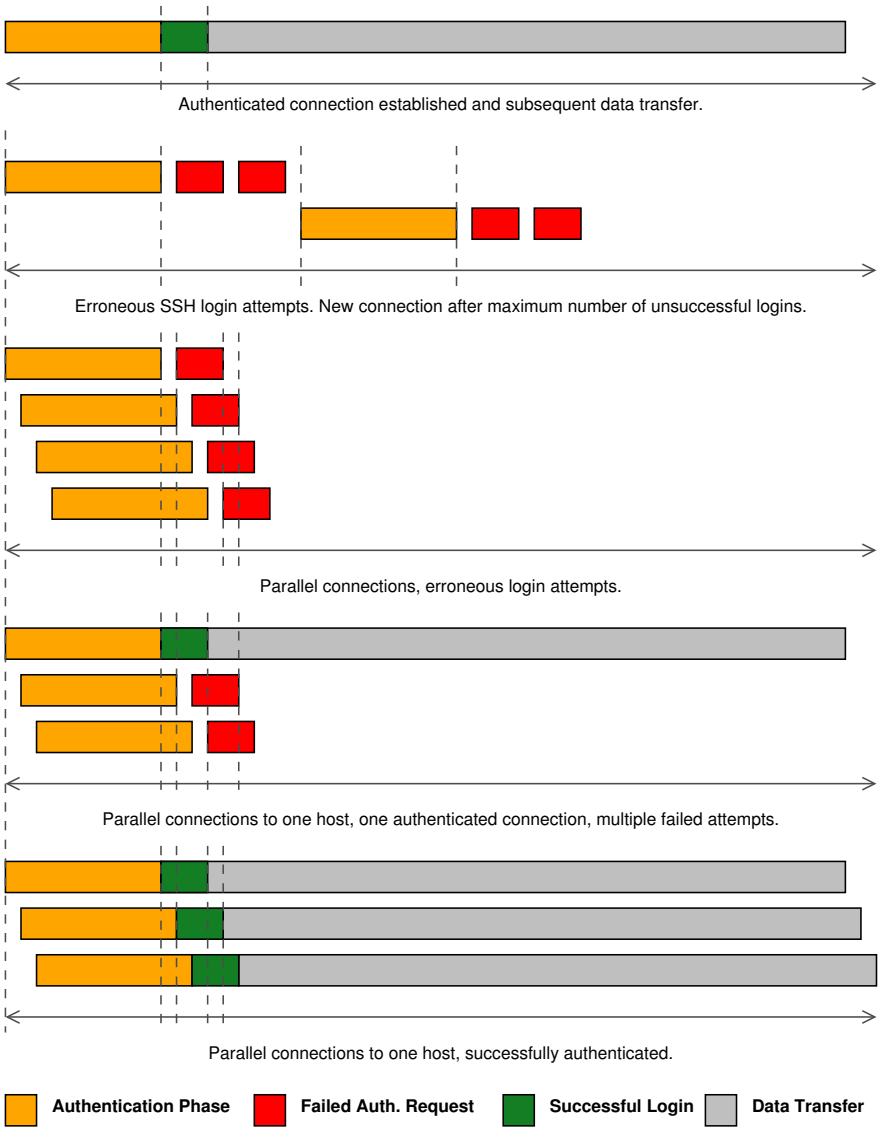


Figure 1. Different cases during the authorization phase of a remote session. After the preamble, the user has to be authenticated, e.g., by a username-password-combination (authentication phase). Based on that, the access is granted (successful login) and data can be transmitted (data transfer) or it is rejected (failed auth. request) and the user has to re-enter the information or the connection is closed.

an unsuccessful try.

- Initiation and successful authentication of a connection, concurrently rejected authentication attempts on parallel connections (e.g., when using NAT).
- Initiation and execution of parallel sessions with subsequent data transfers.

Further cases, e.g., a connection to a port with one unsuccessful authentication try and a subsequent successful one, can be subsumed to the similarity calculations of the given cases.

As indicated in Figure 1, only the sequence of network packets which represent a successful respectively failed authorization attempt is needed for the calculation. Anyway, it is not necessary to know the exact number and position of the involved network packets, because of the subjacent similarity measurement. Therefore, even if some packets are missing or are not belonging to the authentication itself, the similarity values will stay within an evaluable border. By using this approach neither the payload sizes nore the exact number of packets must be known.

The system is based on a transparent network bridge. All the network packets are copied by the use of the `pcap`-library and sent to the detection engine. First, hash-values are built for the identification and management of each connection. Based on the address and port and the payload sizes of the transmitted packets, the used type of connection

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

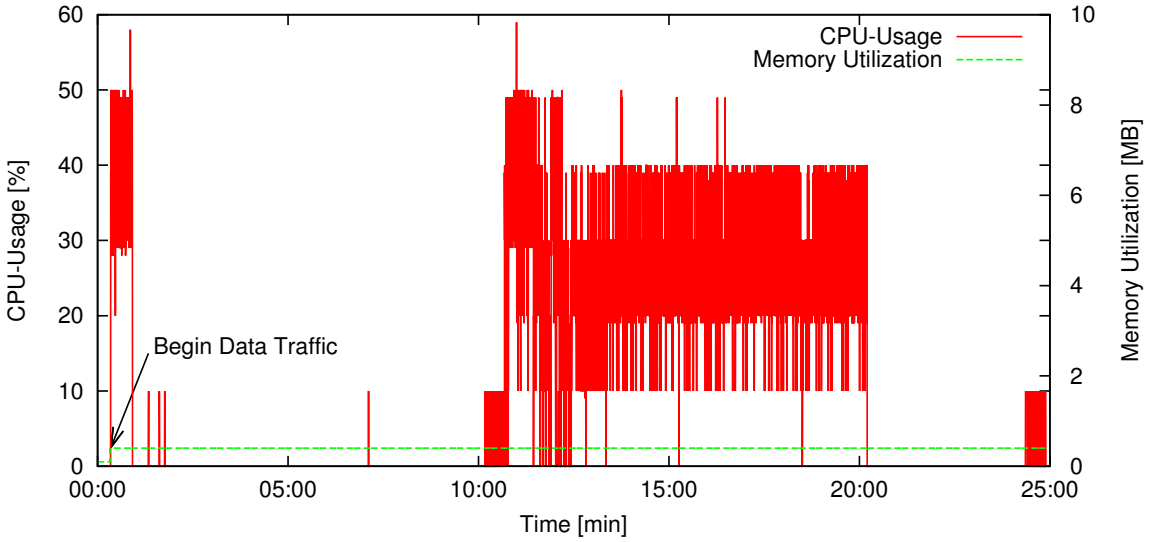


Figure 2. Resource usage of the detection system under load of synthetic traffic.

and correlation is identified, afterwards this information and the packet series are given to the correlation function which calculates their similarity. For the calculation of the similarity, the following formula is used:

$$s_d = \frac{\sum_{i=1}^{n} [(f[i] - m_f) \cdot (g[i-d] - m_g)]}{\sqrt{\sum_{i=1}^{n} (f[i] - m_f)^2} \sqrt{\sum_{i=1}^{n} (g[i-d] - m_g)^2}}$$

m_f and m_g are the arithmetical means of the value series $f[x]$ respectively $g[x]$, d is the considered shifting of $g[x]$. If the values of the two series are similar for a specific d, the standardized cross-correlation will result in a value of one. If the correlation produces zero for all shiftings of d, the series will be uncorrelated. Therefore, the algorithm uses the series of network packets transmitted during the authentication process, searching for areas of high similarity based on repeated rejected authentications. Again, note that it is not necessary to know the exact borders of packets belonging to the authentification process because of the used shifting.

If high similarities are detected a defined number of times, a blocking rule for the `netfilter` firewall is generated and installed on the bridge via a system call of `iptables`. Therefore, the IP address which is the originator of the attack is no langer able to reach the server.

IV. Proof of Concept

For the verification of the efficiency of our approach, a prototype was implemented and tested. The system is realized as a lightweight network-based detector integrated into a transparent bridge. Therefore, it can be placed everywhere into the network, typically close to the border router (therefore being able to analyze all incoming traffic).

Two test methodologies had been used: Firstly, a simulated environment was installed, where both the benign and the malicious traffic was synthetic. Therefore, the behavior and correctness of the system could be decided exactly. Secondly, evaluations in a productive network had been fulfilled, therefore being able to analyze the system under real-world conditions. There, the real traffic of the network presents the background traffic and attacks onto targets in the network had been conducted. By that, the detection system has to operate in the real-word environment and it is possible to evaluate wether all conducted attacks are recognized. Beyond that, it is possible that real attacks appear in the network during the evaluation phase. Because these attacks are not known, the system should detect them, but one can not say if all unknown attacks are detected correctly. To reduce the probability of a wrong evaluation, additional Intrusion Detection Systems were used to analyze the connections on the server. For the execution of attacks, `SSHatter` and `brutessh` had been used as well as manual connections initiated with `ssh`. Different encryption algorithms had been used to verify the robustness of the system. For example, the algorithms `AES128-CBC`, `BLOWFISH-CBC` or `AES128-CTR` can be used by the `Ciphers`-option of *SSH*:

```
ssh -o Ciphers=blowfish-cbc IP-Addr.
```

First, the resource usage during operation on loaded network links was analyzed. The detection system was installed on an Intel P4 computer (3 GHz, 498 MB main memory) which was used for the synthetic as well as the real-world evaluation on a 100 Mbps link.

Figure 2 shows the CPU-usage and memory utilization

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

over a period of 25 minutes of synthetic traffic.

Easy to see, the memory usage rises in the beginning to prepare the initial hashtables but remains below one MB which is constant during the evaluation period. During the evaluation multiple blocks of intense traffic were generated, resulting in higher CPU-load. As one can see, even with the outdated hardware the CPU-usage remains between 20 and 40 percent on average.

The result of the performance evaluation in the real-world environment is shown in Figure 3. As like as in the synthetic case, the CPU-usage remains on 44 percent in average when evaluating a fully-loaded 100 Mbps-link. Also the memory usage for the needed hashtables is quite low and remains on an average of 0.3 MB. Therefore, the system is efficient and able to monitor broadband links with minor hardware resources.

Next, the detection capabilities had been evaluated. An example of the evaluation in the synthetic scenario is shown in Figure 4.

The total number of connections at a time (representing the different remote sessions respectively their authorization phase) are differing based on the simulation an go up to approx. 50 sessions in parallel. Meanwhile, tool-based and manual attacks had been conducted. As soon as an attacking IP was identified, it was blocked and not able to execute attacks any longer for at least the default blocking-time of five minutes. By the evaluation of the detection- and false alarms, a correct classification of about 98.7 percent of all connections was possible. In detail, 92 percent of all conducted attacks had been identified and blocked, while only 0.8 percent of the benign connections had been misclassified as attacks. The false alarm rate of the system is about 0.8 percent while the false alarm ratio (the portion of the alarms which was falsely generated when examine benign traffic) is about 4 percent.

Next, the system performance in a productive network was evaluated. Figure 5 gives an example of a test run over one and a half hour, containing multiple brute-force attacks. It is important to know that in the shown example *all* network traffic passing the transparent bridge was evaluated, not only the traffic of remote sessions. The number of active IPs and used ports are shown as well as the number of concurrently blocked IPs. Because of the readability of the graph, only the beginning of a blocking phase is drawn, not the whole phase or its ending.

In average, about 99.5 percent of all connections had been classified correctly, with a detection probability of about 98.4 percent and a false alarm rate of about 0.5 percent. On the other side, the false alarm ratio of the tests in the productive network is about 51.8 percent. Therefore, half of the generated alarms are based on benign traffic which is quite high. Anyway, this only happens when the *complete* network traffic is examined by the system for the appearance of brute-force attacks, not only the traffic of remote sessions or other login mechanisms. If only the relevant connections are selected and analyzed, the false alarm ratio drops down to about 1.6 percent. This can be achieved by the integration of a protocol analyzer into the system (e.g., see [10], [11], [12]).

V. Related Work

Despite the wide dissemination and broad usage for attacks, only little work is done in the area of brute-force detection. Today's detection mechanisms are typically based on the number of login attempts within a limited time window and originated by a unique IP address. Multiple host-based tools are available which monitor this behavior and block IP addresses, e.g., [13]. There, the evaluation is done by monitoring the logfiles or counting connections made from a single IP address to a configured remote service like *SSH*.

Firewalls can be configured to slow down brute force attacks and block identified IP addresses. For example, the netfilter-firewall can be used to detect brute-force attacks and block IPs by the use of two rules: `iptables -I INPUT -i eth0 -p tcp -m tcp --dport 22 -m state --state NEW -m recent --set --name SSH_BRUTE --rsource` and the second rule `iptables -I INPUT -i eth0 -p tcp -m tcp --dport 22 -m state --state NEW -m recent --update --seconds 300 --hitcount 5 --name SSH --rsource -j DROP`. The first rule dynamically creates a list of IP addresses which is matched as followed: First, the scource address will be added to the list when using port 22. If the address already exists, the entry is updated. Therefore, the already seen addresses can be stored. The second rule monitors the number of connections from this address; if the number of *NEW* connections from the same address and to port 22 exceeds four, the IP address will be blocked for 300 seconds.

The current approaches are easy to use, but have several disadvantages, e.g., only specified ports can be monitored and network performance can be reduced keenly. For example, often the standard port 22 is monitored but *because* of the popular brute-force attacks, many administrators moved the port to non-standard ones. Anyway, attackers often scan their targets, therefore being able to attack non-standard ports, too. Another shortcoming of these approaches is the lack of detecting distributed attacks.

Blacklists are another popular method for avoiding brute-force attacks originated from well-known IP addresses (e.g., [14]). Anyway, because of the intense use of Botnets and therefore an extensive pool of end-user IP addresses, this traditional mechanisms are not adequate any longer. By the distribution of the authentication tries to numerous different IP addresses, a detection based on the number of parallel or consecutive connection tries can be avoided. Also, the use

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

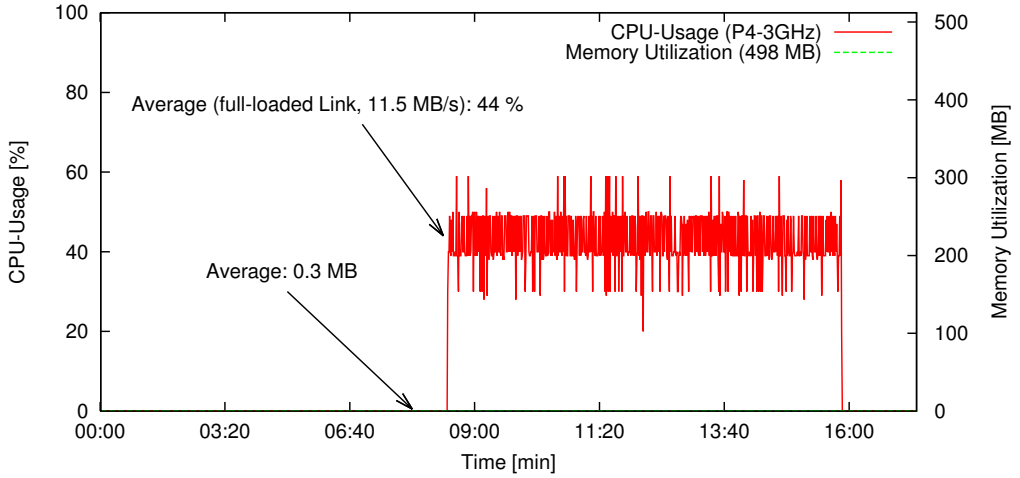


Figure 3. Resource usage of the detection system integrated into a productive network.

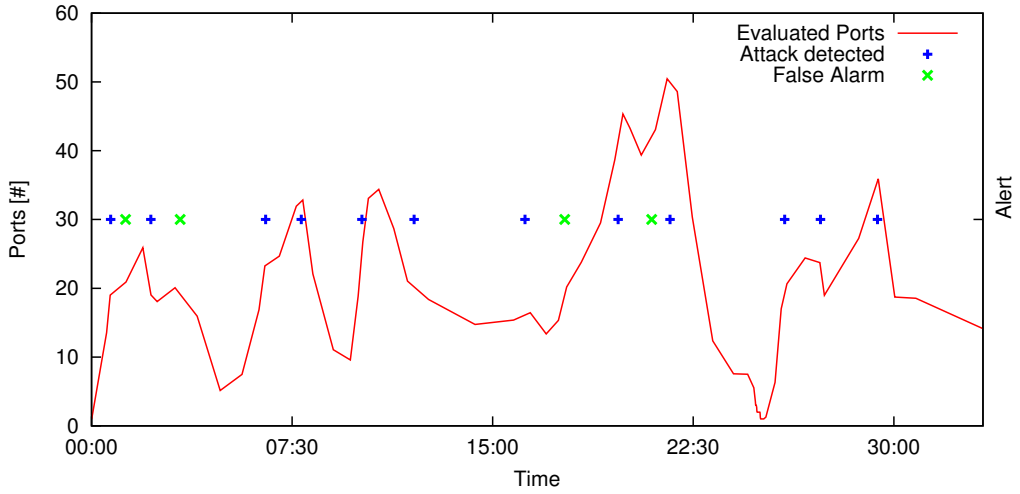


Figure 4. Attack detection and false alarms of the security system in a synthetic environment.

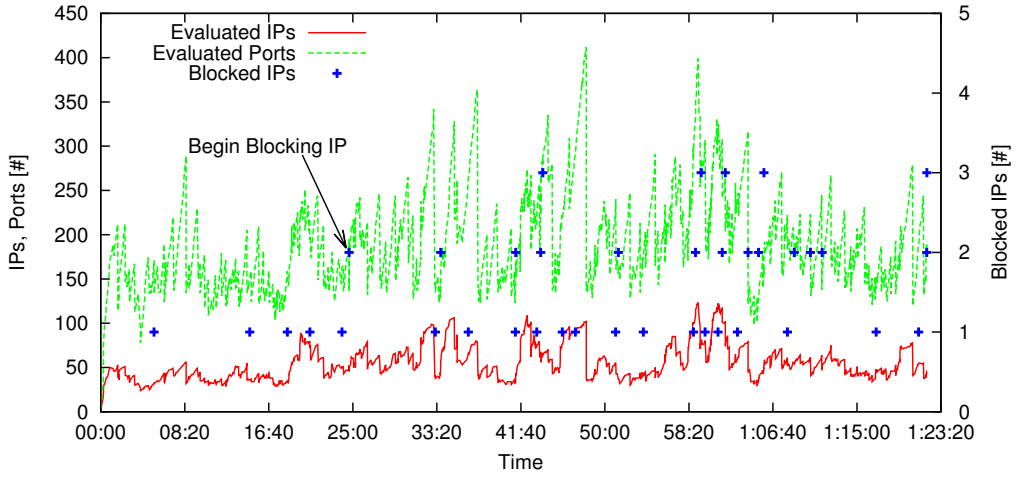


Figure 5. Attack detection of the security system in a productive network.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Table II
DETECTION PROBABILITIES OF THE PROPOSED ARCHITECTURE. VALUES IN BRACKETS: RATIO AFTER FILTERING.

	Classification	Detection Prop.	False Alarm Ratio	False Alarm Rate
Synthetic	98.68	92.0	4.0	0.84
Productive Network	99.53	98.41	51.79 (1.59)	0.48
Combined	99.11	95.21	27.90 (2.80)	0.66

of different idle times, multiple encryption algorithms, etc. can avoid the detection of attacks, too.

VI. CONCLUSION AND FURTHER WORK

Brute-force attacks are a popular and common used attack technique which can greatly endanger systems and networks. Anyway, the current security mechanisms are not able to provide an adequate protection against these attacks.

Therefore, we proposed a new architecture, which is able to provide a fast network-based brute-force detection. Only little system resources are required for the operation of the detection engine. In contrast to existing approaches, the system design enables an attack detection independent of the specific communication between the server and the client, used ports and IPs, the encryption in use, the configuration of the server and the speed and kind of executed brute-force attacks. Therefore, the system can be used in general and without any configuration. All required information is gathered in real-time from the live connections by the use of similarity measurements. It can secure all hosts in a network, disabling the necessity to provide security mechanisms for every single system.

To further improve the detection results and limit the false alarm rates, a pre-filtering of the traffic can be done. At the moment, the complete network traffic is analyzed, but a service detection can be used to analyze only services which are prone to brute-force attacks. By that, the false alarms can be reduced considerably, enabling the system for an application in productive networks. In this connection, we are also going to evaluate the preformance of our system on 1 and 10 Gbps network links.

Also, we are going to adapt the working scheme for a high-efficient Distributed Denial of Service (DDoS) Attack Detection, which is another widespread attack type. By the use of Botnets, DDoS-attacks easily can be used to take down servers and also whole infrastructures of a provider, for example to press money. The presented concepts are the base for an efficient network-based detection and prevention system.

ACKNOWLEDGMENT

This work is done at the Chair for Communication Systems and Internet Services led by Prof. Dr. Dreo Rodosek, part of the Munich Network Management (MNM) Team.

REFERENCES

[1] A. Sperotto, *Flow-Based Intrusion Detection*, PhD thesis, University of Twente, The Netherlands, 2010.

[2] C. Seifert, *Analyzing Malicious SSH Login Attempts*, website, http://www.symantec.com/connect/articles/analyzing-malicious-ssh-login-attempts, last seen on January 19th, 2012.

[3] A. Sperotto, R. Sadre, F. van Vliet and A. Pras, *A Labeled Dataset for Flow-Based Intrusion Detection*, LNCS 5843, pp. 39-50, IPOM 2009, Springer-Verlag Berlin-Heidelberg, 2009.

[4] ATLAS Arbor Networks, *Active Threat Level Analysis System*, website, http://atlas.arbor.net/, last seen on January 19th, 2012.

[5] SSHatter, *SSHatter - Freecode*, website, http://freecode.com/projects/sshatter, last seen on January 22th, 2012.

[6] BruteSSH, *BruteSSH, SSH password brute forcer*, website, http://www.edge-security.com/brutessh.php, last seen on January 22th, 2012.

[7] B. Schneier, *Real-World Passwords*, website, http://www.schneier.com/blog/archives/2006/12/realworld_passw.html, December 2006, last seen January 19th, 2012.

[8] Imperva Application Defense Center, *Consumer Password Worst Practices*, whitepaper, http://www.imperva.com/docs/WP_Consumer_Password_Worst_Practices.pdf, 2011.

[9] C. Herley, *So long, and no thanks for the externalities: the rational rejection of security advice by users*, Proceedings of the 2009 workshop on New security paradigms workshop, pp. 133-144, NSPW 09, ACM, 2009.

[10] L. Bernaille, R. Teixeira, I. Akodkenon, A. Soule and K. Salamatian, *Traffic Classification on the fly*, SIGCOMM Comput. Commun. Rev., 36(2):23-26, April 2006.

[11] A. Moore and K. Papagiannaki, *Toward the Accurate Identification of Network Applications*, In Passive and Active Network Measurement, LNCS 3431, pp. 41-54, Springer-Verlag Berlin-Heidelberg, 2005.

[12] C. Wright, F. Monrose and G. Masson, *On Inferring Application Protocol Behaviors in Encrypted Network Traffic*, J. Mach. Learn. Res., 7:2745-2769, December 2006.

[13] R-FX Networks, website, http://www.rfxn.com/projects/, last seen on January 19th, 2012.

[14] Blacklisting and Abuse Reporting, website, http://www.openbl.org/, last seen on January 19th, 2012.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Correlated M/G/1 Queue modelling of Jitter Buffer in TDMoIP

Usha Rani Seshasayee and Manivasakan Rathinam
Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai-600 036, India.
e-mail: ee08d001, rmani@ee.iitm.ac.in

***Abstract*— Time Division Multiplexing over Internet Protocol (TDMoIP) is a pseudowire technology for emulating TDM circuits over Internet Protocol (IP) networks. In such networks, timing and synchronization plays a key role in achieving the required jitter in terms of variance of interdeparture interval. A jitter buffer is used at the receiver, to circumvent the impairment of the packet networks: delay, jitter and loss. But, out of these, delay and loss can't be compensated for, while QoS in IP networks is used to minimize them. The jitter (or variance of packet delay) can be reduced to a tolerable level at the receiving Inter Working Function. A tradeoff between delay and jitter is required to achieve the desired jitter. This paper presents the condition under which the jitter buffer at the receiver is to be operated for minimum output variance in a TDMoIP framework, to achieve minimum slip rate and thus better voice quality. The receiver jitter buffer is modeled as a correlated M/G/1 queueing system with EARMA correlations between the interarrival and the service times. The motivation for the above correlation structure is that, given the correlations within the service intervals, the EARMA correlation results in reduction of variance in the interdeparture interval. This is a step towards achieving CBR upstream. The key advantage of using EARMA correlation is that the analysis of such a correlated queue is analytically tractable. The variance of the interdeparture times of the above queue is presented. The analysis of the departure process, the waiting times of incoming packets of this correlated queue and the relevant simulations show that if the variance of the interdeparture time process constituting output TDM stream is to be less than that of the interarrival time process of the jitter buffer, which is modeled as M/G/1 queue, then the mean waiting time of the packets in the jitter buffer would be greater than that of independent (M/M/1) case. The values of the parameters of the M/G/1 queue which minimizes the variance of interdeparture interval are identified. Our study also included a G/G/1 queue in which interarrivals are also correlated. Extensive simulations demonstrate our analytical results.**

Keywords-TDMoIP; Jitter; Correlated queue.

I. INTRODUCTION

Time Division Multiplexing (TDM) circuits have been the backbone of voice communications over the past several decades. TDM is a reliable, hard partitioned circuit switched technology and offers low delay services for real time interactive digital telephony. But, the bandwidth is used inefficiently in TDM. For efficient bandwidth utilization, there has evolved a 'converged' network catering to all services that is packet based. In such a network, digitized voice is carried over packet switched infrastructure. The transition from circuit switched to packet switched would take place in a phased manner, as legacy networks could not be replaced overnight. Industries and academia are trying to provide solution to this as TDM over Packet Switched Network infrastructure. In this work, we address synchronization in Time Division Multiplexing over Internet Protocol (TDMoIP).

Time Division Multiplexing over Internet Protocol [1, 2, 3] is a technology wherein a logical circuit is realized in an IP network, which links two TDM islands. The TDM traffic at the transmitter is packetized into constant bit sized frames and transmitted across an IP network. When the packet carrying the TDM payload traverses the IP network, it experiences random delay due to the queueing at the intermediate routers. It is because of this, that the packets at the receiver arrive randomly, an effect called as packet delay variation/jitter. To compensate for this, jitter buffer of large enough size is used. A mismatch between the read and the write clocks at the input and the output of this buffer, due to large delay variation will cause overflow or underflow of the jitter buffer. Such clock mismatch can lead to observable defects on the end service, e.g. frame slips when interworking with existing PSTN and narrowband ISDN (N-ISDN) networks [2]. Synchronization in the data link layer of the ISO stack is therefore an important issue in such networks.

The area of timing and synchronization is well researched. Clock recovery schemes can be categorized into synchronous and asynchronous. In case of synchronous schemes [4], the sender delivers via the outgoing stream, the information on the frequency difference between a source clock and reference clock, which is generated out of a common clock available to both sender and the receiver. The original source clock is recovered using this frequency difference. Asynchronous schemes recover the source clock from one of the following: buffer level [5], time difference of arrival of packets [2]. Hybrid techniques which combine any

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

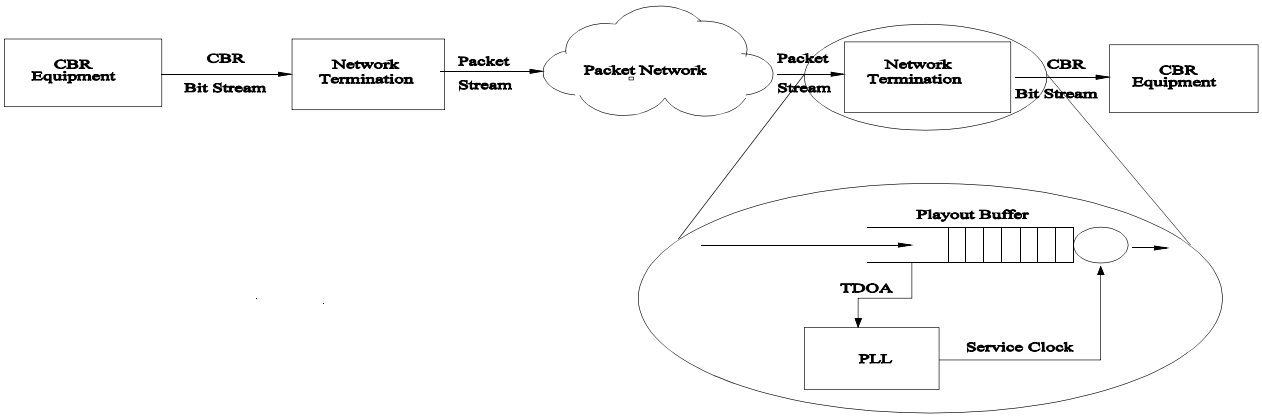


Figure 1. A conventional model for clock recovery for TDMoIP

of two methods have also been studied [6]. All of the above schemes attempt to recover the source clock in the physical layer. For achieving synchronization at the data link layer of the ISO stack, in this paper, we use [7] a queueing model, where frames arriving from the transmitter are queued in the jitter buffer and served such that the variance of the interdeparture process of the outgoing frame stream is minimum.

The rest of the paper is organized as follows. Section 2 briefly describes conventional and proposed model for TDMoIP. Section 3 discusses the proposed model in detail. An analytical expression for variance of the departure process at the receiver is also presented. We are the first one to look at the departure process of a correlated queue, to characterize jitter in packet networks. Section 4 gives the simulation results for interarrival times being independent as well as, correlated. In Section 5, conclusions are drawn and future work is outlined.

II. CONVENTIONAL MODEL FOR TDMOIP

TDMoIP is a technology wherein TDM frames such as E1/T1 are packetized and sent across a layer 2/3 virtual circuit, in which the QoS is provisioned such that these TDM encapsulated packets are given the highest priority. These packets upon traversing the packet network experiences delay variability, which is caused by the queueing delays at the intermediate routers. These packets having the TDM payload, on arriving at the receiver are stored in a jitter buffer, which is meant to mitigate the effect of the packet network.

A. Schematic for TDMoIP

The scheme for clock recovery at the physical layer [2] is shown in Fig. 1. For TDMoIP, there is no common clock present at the transmitter and receiver. The receiver has to estimate the transmitter clock from the received data stream. This is accomplished using a phase-locked loop (PLL). The PLL locks the phase of the receiver clock to that of the transmitter or would manipulate the arrival pattern to get accurate receiver clock. Thus the function of the PLL is to compensate for any frequency deviation. The main problem encountered in packet networks is the queueing delay (which is variable for each packet), introduced by packet routers and switches. It is because of this varying delay that the constant bit rate TDM stream does not appear periodic. A frequency estimate of the transmitter clock is done based on the measurements of the packet arrivals to mitigate the effects of packet delay variations using various techniques. This frequency estimate is used as reference for the PLL. The PLL locks the receiver clock's phase to that of the transmitter. The clock from the PLL is used to derive the Constant Bit Rate (CBR) bit stream. Therefore, the better the estimate of frequency, the more periodic will the received stream be. The variance of the frequency estimate should be low for proper synchronization.

B. Literature survey

When there is no common network clock, the asynchronous clock recovery schemes are applicable. They are also called as adaptive clock recovery schemes. An adaptive clock recovery method based on jitter buffer level was studied in [5]. In this jitter buffer level based method, the clock is recovered by calculating the receiving rate of the IP packets. In another adaptive scheme, the time difference of arrivals of the packets is filtered and source clock is recovered with minimum variance as in [2]. There are hybrid techniques, which combine two of the above methods to minimize the error in clock recovery and also to increase its convergence rate [6]. The schemes mentioned here recover clock at the physical layer. Minimizing variance of the inter-packet time at either of the two layers, physical or data link of the ISO stack will reduce jitter in the outgoing stream.

For achieving synchronization at the data link layer of the ISO stack, in this paper we use [7] a queueing model, where payloads are extracted from the TDM encapsulated IP packets arriving from the transmitter, are queued in the jitter

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

buffer. They are served such that the variance of the inter-departure process of the outgoing packet stream is minimum.

C. *The Jitter buffer model*

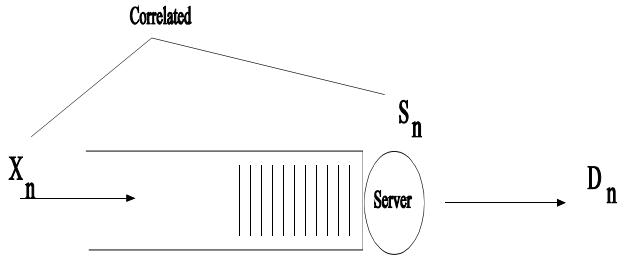


Figure 2. FIFO, Single server queue

The receiver jitter buffer is modeled as a single server, first in first out (FIFO) queue where $\{X_n\}$ is the interarrival time process, $\{S_n\}$ is the service time process, with correlation between the two processes, as shown in Fig. 2. The interdeparture time process of this queue should have minimum variance so that there is proper synchronization at the receiver.

Jacobs [7] gives the waiting time distribution for a correlated queue having correlations between interarrival and interdeparture times under heavy traffic conditions. We use the above model for modelling the jitter buffer and study the interdeparture process in such a queue. We are the first one to look at departure process, its statistics and to use the same in a jitter buffer model. For the analysis of the simulation data of the queue [8] is used. We include here in our studies a queue with non-Poissonian arrival process, to be more realistic.

III. OUR QUEUE MODEL

A. *Research solution: The EARMA queue model*

A M/G/1 queue is a queueing system with a single server, arrival process being Poisson, service times following a general distribution and having an infinite buffer space. The jitter buffer is modeled as a M/G/1 queue with exponential autoregressive and moving average (EARMA) [7] correlations between the interarrival times and the service times. The following table gives the notations used for the correlated queue.

TABLE I. THE EARMA QUEUE

Symbols	Description
$\{X_n\}$	Sequence of independent exponentially distributed random variables with positive finite mean λ^{-1}, where X_n is the interarrival time between the n th and the n-1 th arrival of the packet at the receiver jitter buffer
$\{E_n\}$	Sequence of independent exponential random variables with positive finite mean μ^{-1}
$\{J_n\}$	Sequence of Bernoulli random variables with $P(J_n=1)=1-\beta$; $0\le\beta\le1$
$\{K_n\}$	Sequence of Bernoulli random variables with $P(K_n=1)=1-\rho_{ar}$; $0\le\rho_{ar}<1$
$\{S_n\}$	The service time process, where S_n represents the service time of the n th packet, following a general distribution and having a positive finite mean μ^{-1}
$\{B_n\}$	An auto regressive process, where B_0 has an exponential distribution with mean λ^{-1}

We consider a single server queue with FIFO discipline, as the packets that are sent from a single source arrive in the order transmitted inspite of the random delay. The 0^{th} packet arrives at t=0 and finds the server free. Packets are of constant size, an E1 frame per packet. Henceforth, the word packet goes synonymous with the word frame and treated as an entity arriving at the queue.

In order to reduce the variance of the departure process, we introduce a positive correlation between service intervals and interarrival times (X and Y are said to be positively correlated when X increases Y also increases). This would mean that a queue with large buffer occupancy would be served faster and vice versa. All these would reduce variance of interdeparture times. A correlation structure which satisfies our requirement is the EARMA correlation. Hence, the service time of the n th packet, S_n is taken to be:

$$S_n = \beta E_n + J_n(\lambda\mu^{-1}B_n) \tag{1}$$

where, $B_n = \rho_{ar}B_{n-1} + K_nX_n$ (2)

Also, note that statistical correlation is also imposed within the service times themselves. These models have the advantage that the marginal distributions and correlation structure of the sequences are specified separately, so that we can compare them with an independent (M/M/1) case with the same distribution. The covariance within service times is obtained from (1) and (2) as

$$COV(S_n, S_{n+k}) = \mu^{-2}\rho_{ar}^k(1-\beta)^2; k = 1,2,3..... \tag{3}$$

The covariance between the interarrival and service times is obtained from (1) and (2) as

$$COV(S_n, X_{n-k}) = (\lambda\mu)^{-1}\rho_{ar}^k(1-\rho_{ar})(1-\beta); k = 0,1,2,3....., n-1 \tag{4}$$

The variance of the interdeparture time process should be as less as possible so that the output data stream containing the TDM payload would be closer to CBR as was sent by the transmitter. Larger variance in the inter-departure times would lead to underrun or overrun of the jitter buffer in the upstream node causing clock slips. In effect, all these would result in low quality of voice signal carried by the TDM pipe. By using more sophisticated clock recovery algorithms, recovered TDM clocks can be made to comply with ITU-T G.823 and G.824 specifications for T1/E1 jitter and wander control while simultaneously delivering optimal latency.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

B. *Departure process*

Departure process is important in the analysis of TDMoIP synchronization algorithm. The variance of the interdeparture time should be minimum as possible so that the output stream would appear periodic with minimum packet delay variation. Table II gives the variables related to the departure process.

TABLE II. DEPARTURE PROCESS OF THE QUEUE

Symbols	Description
$\{D_n\}$	Sequence of random variables with positive finite mean λ^{-1}, where D_n is the interdeparture time between the n th and the n-1 th packet at the output of the receiver jitter buffer
$\{W_n\}$	The waiting time process, where W_n is the waiting time of the nth packet in the Queue

From the queueing theory results [9, 10, 11], the inter-departure time process can be written in terms of the interarrival, waiting and service times. Let $\{D_n\}$ be the inter departure time process, where D_n is the interdeparture between the n th and the n-1 th departure of packets at the output of the receiver jitter buffer, then

$$D_n = X_n + W_n - W_{n-1} + S_n - S_{n-1} \tag{5}$$

From the above equation, the variance of the interdeparture time process is written in terms of the variance and covariance terms as given below. Other covariance terms which are not present in this equation tend to zero because the two variables in them are independent. So we have,

$$\begin{aligned} Var(D_n) &= Var(X_n) + 2.Var(S_n) + 2.COV(S_n, X_n) - 2.COV(S_n, S_{n-1}) \\ &\quad - 2.\mathrm{E}[X_n - S_n].\mathrm{E}[W_n] \\ &= \lambda^{-2} + 2\mu^{-2} + 2(1-\rho_{ar})(1-\beta)(\lambda\mu)^{-1} - 2\mu^{-2}\rho_{ar}(1-\beta)^2 \\ &\quad - 2[\lambda^{-1} - \mu^{-1}].\mathrm{E}[W_n] \end{aligned} \tag{6}$$

The variance of Poisson arrival process is λ^{-2}.Therefore, the condition for the variance of inter-departure time process, which is stationary, to be less than the variance of the interarrival time process from (6) is obtained as:

$$\mathrm{E}[W_n] > \frac{\lambda}{\mu(\mu-\lambda)}\left[1 + (1-\rho_{ar})(1-\beta)\mu\lambda^{-1} - \rho_{ar}(1-\beta)^2\right] \tag{7}$$

Equation (7) is satisfied only when the parameters are related as: $\rho_{ar} > (2-\beta)^{-1}$, especially under heavy traffic limit conditions. But under such a condition, the mean waiting time would be definitely be greater than the independent case. Thus, there is a tradeoff between the mean waiting time and the variance of the interdeparture time process. The study of the variance of the departure process is done more precisely here, than in [12].

C. *Correlated interarrivals*

The queue was studied with interarrival times being correlated as in [13], as it was found in [14], that the interarrivals are correlated in packet networks. The jitter buffer is modeled as a G/G/1 queue, with correlation within the interarrrival times. Table III gives the variables related to the correlated arrival process.

TABLE III. CORRELATED INTERARRIVAL TIMES

Symbols	Description
$\{G_n\}$	Sequence of independent exponentially distributed random variables with positive finite mean λ^{-1}
$\{H_n\}$	Sequence of Bernoulli random variables with $P(H_n=1)=1-\alpha$; $0 \le \alpha < 1$

The interarrival time between the n-1 th and the n th arrival, X_n is given by:

$$X_n = \alpha\, X_{n-1} + G_n H_n \tag{8}$$

IV. SIMULATION RESULTS

The correlated M/G/1 queue with EARMA correlations was simulated for different traffic intensity and correlation parameter values in MATLAB. The graph for the ratio of mean waiting time of this correlated queue to the mean waiting time of the independence (M/M/1) case was plotted. Also, the ratio of the variance of the departure process to the arrival process was plotted, in the same graph. The simulation results are interpreted clearly than in [12].

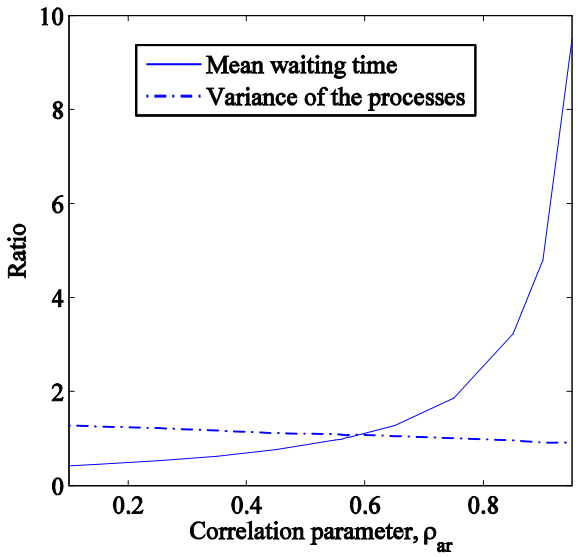


Figure 3. Ratio of the mean waiting time of the correlated M/G/1 to the M/M/1 queue and ratio of the variance of the departure process to the arrival process, for various correlation parameter values, ρ_{ar}; β=0.25; traffic intensity, ρ=0.9 Erlangs.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

As seen above, under heavy traffic condition, the ratio of the variance of departure process to the arrival process of this queue decreases when strong correlation is imposed between the interarrival and service times and goes well below a factor of 1. But the ratio of the waiting time of this queue to the independent case increases when there is strong correlation between the interarrival and service times. The correlation parameter, $\rho_{ar}= (2-\beta)^{-1}=0.57$, when $\beta=0.25$ is the point from where the variance of the departure is less than that of the arrival process. Therefore, as seen from Fig. 1, high correlation parameter values yields lesser variance, but the waiting time is more.

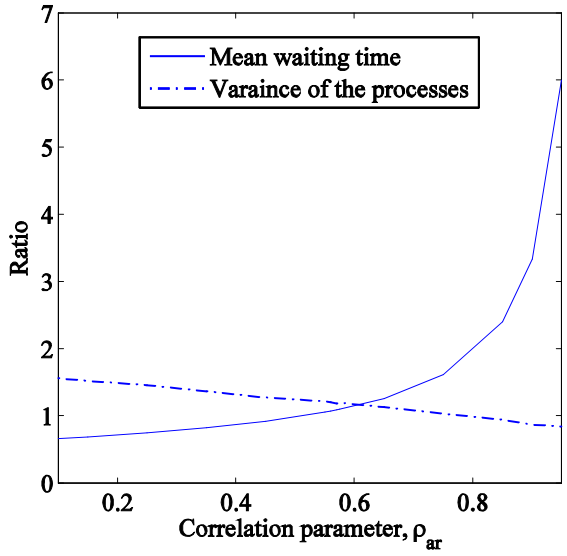


Figure 4. Ratio of the mean waiting time of the correlated M/G/1 to the M/M/1 queue and ratio of the variance of the departure process to the arrival process, for various correlation parameter values, ρ_{ar}; $\beta=0.25$; traffic intensity, $\rho=0.6$ Erlangs.

As is seen from Fig. 4, even under light traffic condition, the ratio the variance of the departure process is less than arrival process for strong correlation, that is, higher values of ρ_{ar}. The correlation parameter values at which the mean waiting time ratio and the variance ratio are unity are same, as for the heavy traffic condition as in Fig. 3. That is, the correlation parameter, $\rho_{ar}= (2-\beta)^{-1}=0.57$, when $\beta=0.25$ is the point from where the variance of the departure is less than that of the arrival process.

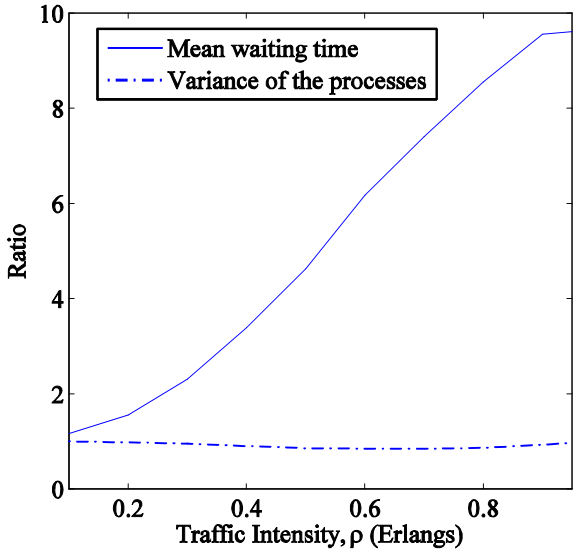


Figure 5. Ratio of the mean waiting time of the correlated M/G/1 to the M/M/1 queue and ratio of the variance of the departure process to the arrival process for various traffic intensity values, ρ (Erlangs); $\beta=0.25$; correlation parameter, $\rho_{ar}=0.95$.

Fig. 5 depicts a condition where the correlation parameter is selected such that the variance of the departure process is less compared to the arrival process but the mean waiting time is more compared to the independent case. That is, our objective of minimum jitter variance is achieved.

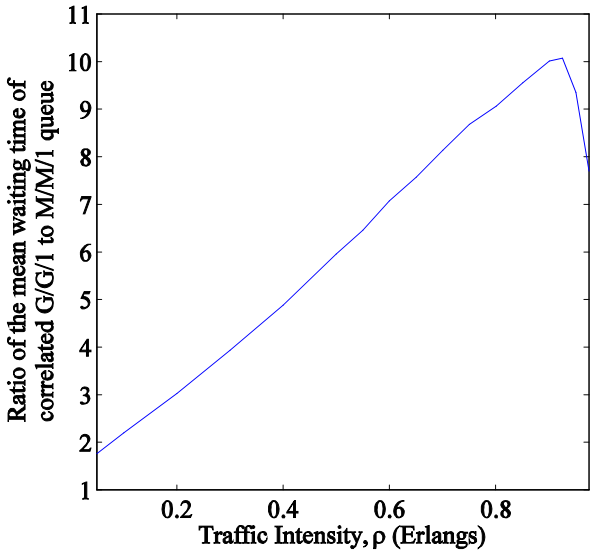


Figure 6. Ratio of the mean waiting time of the correlated G/G/1 to the M/M/1 queue for various traffic intensity values, ρ (Erlangs); $\beta=0.25$; correlation parameter, $\rho_{ar}=0.95$; interarrival time correlation parameter, $\alpha=0.65$.

Fig. 6 depicts a condition where the correlation parameters β and ρ_{ar} are selected for a particular interarrival time correlation parameter, α such that the mean waiting time is more compared to the independent case. High mean waiting time renders low variance of the departure process. This aids in achieving lesser jitter. Under heavy traffic conditions, the ratio of the mean waiting time of the correlated G/G/1 to the M/M/1 queue reduces. For this particular value of the interarrival time correlation parameter, it is better not to operate the queue under heavy traffic condition, so as to achieve our objective of lesser variance.

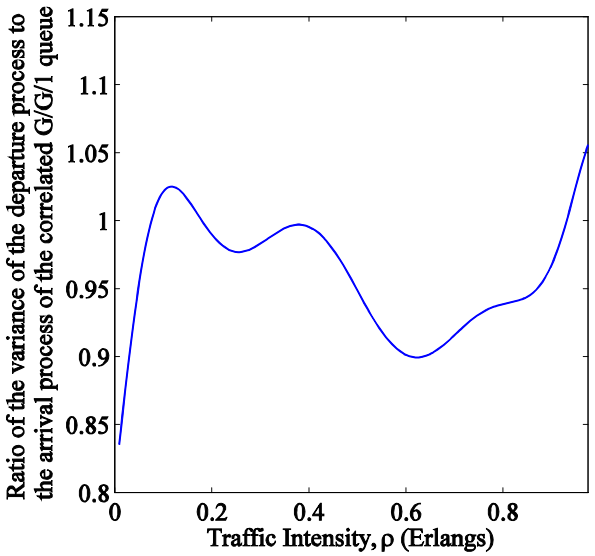


Figure 7. Ratio of the variance of the departure process to the arrival process for various traffic intensity values, ρ (Erlangs); β=0.25; correlation parameter, ρ_{ar}=0.95; interarrival time correlation parameter, α=0.65.

Fig. 7 depicts a condition where the correlation parameters are selected such that the variance of the departure process is less than the independent case, for a particular value of the interarrival time correlation parameter.

The queue behaviour, when there is auto-regressive correlation (correlation value, α=0.65) between interarrival times shows that the mean waiting time of this correlated queue is greater than the mean waiting time of the M/M/1 queue at a correlation parameter value, between the interarrival and the service time, which is different from the one presented earlier. Whereas, the variance of the departure process of this queue is less than the variance of the arrival process. Only when the correlation between the interarrival times is less, the objective of lesser variance of the departure sequence is achieved.

V. CONCLUSION AND FUTURE WORK

A correlated M/G/1 is used for modelling the jitter buffer in TDMoIP and a set of parameters of the queue is identified to achieve our objective of minimum variance of the departure process. The correlation parameters are obtained to achieve the desired output variance. This work is extended to arrivals being correlated, as interarrival times are correlated in packet networks and their simulation results are presented. Obtaining the analytical expressions relating the correlation parameter of the interarrival time and the queue statistics can be a future work. A feedback mechanism to control the various correlation parameters to achieve the desired voice quality can also be considered.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] H. M. Ahmed, “Adaptive Terminal Synchronization in Packet Data Networks,” IEEE Globecom, pp. 728-732, 1989.

[2] J. Aweya, D. Y. Montuno, M. Ouellette and K. Felske, “ Clock recovery based on packet inter-arrival time averaging,” Comp. Comm., vol. 29, pp. 1696–1709, 2006.

[3] Y. J. Stein, “TDM Timing,” RAD Data Comm., August 2006.

[4] R. C. Lau and P. E. Fleischer, “Synchronous Techniques for Timing Recovery in BISDN,” IEEE Trans. Comm., vol. 43, pp. 1810–1813, Feb. 1995.

[5] R. P. Singh and S. H. Lee, “Adaptive Clock Synchronization Schemes for real-time traffic in broadband packet networks,” EUROCON 88, pp. 84-88, Jun. 1988.

[6] S. Zhu and Y. Xu, “The Study and Analysis of Joint Adaptive Clock Recovery Mechanism for TDMoIP,” ICNSC, pp. 533-538, Apr. 2008.

[7] P. A. Jacobs, “Heavy traffic results for single-server queues with dependent (EARMA) service and interarrival times, ” Adv. Appl. Prob, Ireland, vol. 12, pp. 517-529, 1980.

[8] A. M. Law, “Statistical analysis of the output data,” Oper. Res., vol. 31, no. 6, pp. 983-1029, Nov. 1983.

[9] L. Kleinrock, “Queueing Systems,” vol. 1, John Wil. & Sons., 1975.

[10] J. W. Cohen, “The Single Server Queue,” vol. 8, North-Holland, 1982.

[11] J. Medhi, ”Stochastic Models in Queueing theory,” second edition, Acad. Press, 2003.

[12] S. Usha Rani and R. Manivasakan, “On the Departure Process of Jitter Buffer in TDMoIP,” Proceedings of the 18th Nat. Conf. on Comm., February 2012.

[13] R. Manivasakan, U. B. Desai and A. Karandikar, “Broadband Teletraffic Characterization using Correlated Interarrival time Poisson Process (CIPP),” J. Ind. Inst. Sci., vol. 79, no. 3, pp. 233-249, 1999.

[14] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, “On the self-similar nature of Ethernet Traffic,” ACM SIGCOMM, Comp. Comm. Rev., vol. 25, pp. 202-213, 1995.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

A Didactic Platform for Testing and Developing Routing Protocols

Adam Kaliszan
Chair of Communication and Computer Networks
Poznan University of Technology
ul. Polanka 3, 60-965 Poznań, Poland
Email: adam.kaliszan@gmail.com

Mariusz Głąbowski
Chair of Communication and Computer Networks
Poznan University of Technology
ul. Polanka 3, 60-965 Poznań, Poland
Email: mariusz.glabowski@put.poznan.pl

Sławomir Hanczewski
Chair of Communication and Computer Networks
Poznan University of Technology
ul. Polanka 3, 60-965 Poznań, Poland
Email: slawomir@hanczewski.pl

***Abstract*—This paper presents a platform for testing and the development of new routing protocols. The platform is an alternative to already existing solutions of the type based on physical devices or on virtualization. In the proposed solution, the testbed nodes are simple routers, i.e., the devices of System on Chip type, with embedded Linux system. These routers perform only packet switching functions, i.e., the function of the Data Plane. The functions related to supporting routing protocols, i.e., the functions of the Control Plane, for all nodes have been moved to a dedicated computer. The Control Plane functions are provided by the Quagga software router, modified for the purposes of the platform. With low cost and small size of single nodes, the platform can also be used in teaching.**

Keywords-Routing protocols; Software Router.

I. INTRODUCTION

The computer networks classes, conducted in Chair of Communication and Computer Network at Poznan University of Technology, among others, are usually carried out using proprietary devices such as Cisco, Juniper or Allied Telesis. The advantage of lab classes with the use of professional routers and switches is to enable students to get familiarized with a configuration of devices that they might meet in practice – in corporate and providers' networks. The disadvantage of this solution that was reported by students, is the operating systems proprietary code for these devices, and hence no chance of modifications and testing of networks protocols, i.a. routing protocols.

The following two solutions that allow for testing of routing protocols have been used in the hitherto known platforms:

- Hardware, i.e., the one where each testing network's node is an independent, totally functional router (either a router or a PC computer with an appropriate software),
- Virtualization, which enables to build a test network of a particular typology on a single physical server.

The solutions based on physical nodes are characterized by high stability — each node is independent and its load does not decrease performance of other nodes. Such a network is, however, difficult to maintain, especially as far as teaching laboratories are concerned where various subjects are being held/lectured (the need to manage connections of different network topologies). Depending on the type of equipment used, the costs of building such a network can be large and, as previously indicated, the possibility of modifying the protocols in case of proprietary solutions, can be significantly reduced or even impossible. Therefore, the testbed of this type is generally built on the basis of PCs running under Linux. For the implementation of new protocols, it is necessary to update the software on each node. An update procedure itself is sometimes very time-consuming.

In the other existing solution used for testing routing protocols, testbeds are using topology virtualization techniques, i.e., they are made of virtual machines (that are a network's nodes), embedded on a single physical server (working under the Linux system). The virtual testbed allows for setting up any connection topology for an indefinite period of time, a quick network's reconfiguration and easy changes in the number of nodes (their number depends on the server performance).

The main cost of building a testbed in the mentioned solution is a purchase of the server. Despite the undoubted advantages of this solution, the obtained test results, such as routing protocols output, might not be reliable, because an overload of a single node may have a negative impact on performance of other nodes, which results from task division of the server's processors.

In order to find the best way of how to eliminate the drawbacks of hardware and virtualization solutions, a new concept of the routers' architecture has been developed in the Chair of Communications and Computer Networks at Poznan University of Technology. It allows students to conduct advanced tests (along with possible modifications) for existing routing protocols, as well as to start and test newly developed routing protocols within class hours (including thesis). The proposed solution combines both the advantages of platforms that use physical devices only with those offered by virtualization. The nodes in a proposed testbed are built from very simple devices, responsible only for switching of the packets (Data Plane). This is the System of the Chip devices, running under Linux. As a result of

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

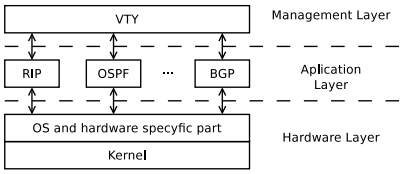


Figure 1. A general architecture of the software router

such an approach, each individual network node is physically independent, and the cost per unit does not exceed 20 euros. The functions responsible for routing (Control Plane) have been transferred in the suggested architecture to a dedicated computer that is running the Quagga software router. The handling of the Control Plane functions of all nodes is implemented by Quagga, after its appropriate modification. The independence of routing function for given nodes is obtained by activating a separate routing protocol process for each of them. The architecture obtained in this way is characterized by high simplicity and low building cost. With full access to the existing routing protocols and the ability to run new protocols, the proposed solution perfectly fits in the teaching of computer networks.

The further part of the article is structured in the following way. Section II presents the idea of software routing and gives an overview of popular solutions of this kind. Section III describes the concept of teaching networks, proposed in the article. Section IV includes a description of implementation procedures. Section V shows the usage scenarios of the elaborated platform. Section VI is a summary of the article.

II. Software router

At present, the most popular router projects with an open source are: Quagga [1] [2] and Xorp [3]. Figure 1 presents the general architecture of the software router. It consists of three layers: the hardware, application and the management layer.

The hardware layer uses the API operating system, or it refers directly to the hardware resources. This leads to its dependence on the operating system and on the hardware platform. The hardware layer is responsible for the preparation of the routing table based on the information received from the application layer. The selection of the route for a packet and its switching to the certain output port is possible owing to the entries in the routing table. The packet switching process takes place in the Data Plane (DP). Packet switching can be implemented via hardware or software. The hardware layer collects information on the status of the interfaces and on their Data Link layer addresses and Network layer addresses. The information on the status of the interfaces and their addresses is then passed on to the application layer, i.e., to the process supporting a given routing protocol.

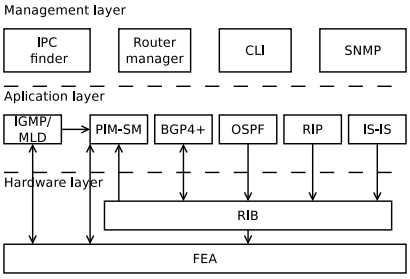


Figure 2. The architecture of XORP router

The application layer communicates with the hardware layer (kernel sublayer) via a special interface. This interface provides a hardware abstraction, which makes the application layer independent of the hardware platform and the operating system. Information on routing paths is sent to the hardware layer using the defined interface (between the application layer and the hardware layer). The applied solution allows simultaneous functioning of multiple processes in the application layer that are associated with various routing protocols. At the same time, the support of many routing processes does not reduce the system stability, since each process runs independently of the others. This solution also makes an easy addition of a new process with new routing protocol possible.

The management layer simplifies the configuration of the routing protocols. It provides access to a configuration of all protocols by CLI (Command Line Interpreter) or other protocols, e.g., WWW (World Wide Web), SNMP (Simple Network Management Protocol), TL1 (Transaction Language 1), etc.

The applied layered router architecture makes it easier to transfer the software of a router onto another operating system/platform, since the required modifications are mainly related to the hardware layer. This vastly simplifies the routing protocols migration between the devices and enables an easy and fast starting of software routers on many hardware platforms running on different operating systems.

The XORP project opts for the convenience of implementation, obtained thanks to C++ language. The whole code has been very well documented [4]. Figure 2 presents the architecture and the functional division of XORP router into particular modules. A possibility of multicast support has been additionally provided in the project. An implementation of the XORP project in the C++ programming language is less efficient; hence producers of network devices, such as software routers, choose the Quagga project.

The Quagga project is a fork of the Zebra software router project. A particular emphasis in Quagga project was put on the productivity and, therefore, the whole Quagga code was written in the C programming language. Rather than using the standard libraries, new ones have been written specially for the project's needs in order to achieve the

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

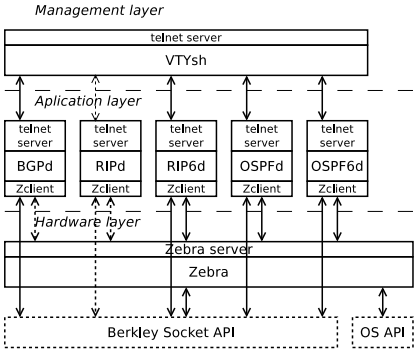


Figure 3. The architecture of Quagga router

Bits 0-15	Bits 16-23	Bits 24-31	Bits 32-47	Bits 48 - ...
Message size	Marker	Version	Command	Message data

Figure 4. The message format of the Zebra protocol

highest performance. Figure 3 presents the architecture of a Quagga router. The Zebra module is responsible for the API operating system support. It reads status of the interfaces and their addresses. It also modifies the system routing table. The Zebra module supports the following operating systems: Linux, Solaris, FreeBSD.

The modules of the application layer (ospf6d, ospfd, rip6d, pipd, bgpd) connect with the Zebra module via a local or TCP connection. The managing of modules in the application layer is possible by using CLI. Each module has a telnet server and supports multiple CLI sessions simultaneously. The VTYsh management layer module connects to all modules of the application layer. At the same time, it provides a telnet protocol server to which users can connect as well. Owing to the VTYsh module, the user has access to all modules of the application layer from a single console that supports the connection with VTYsh. It should be noted that the management layer in the Quagga project is not indispensable. Its addition is to make the command interpreter similar to the one used on Cisco routers.

The following section presents the idea of the modification of the Quagga project. It will consist of adding some functionality to the interface between the hardware layer and the application layer. The interface between the hardware layer and the application layer is described further in this section. The Zebra module operates as a server to which clients are connected – the application layer modules. The communication between the modules is provided by the Zebra protocol. This protocol does not have any documentation and was changing with the evolution of the project. Figure 4 shows the message format of the Zebra protocol. The first field **message size** specifies in bytes the whole message size (along with the header). It is a 16-bit field and the bytes are written in network order. Next 8-bit field **marker** is introduced to keep the compatibility with an older version

Table I
ZEBRA PROTOCOL MESSAGES

code	command	dir
1	ZEBRA_INTERFACE_ADD	C $\leftrightarrow$ S
2	ZEBRA_INTERFACE_DELETE	C $\leftrightarrow$ S
3	ZEBRA_INTERFACE_ADDRESS_ADD	C $\leftarrow$ S
4	ZEBRA_INTERFACE_ADDRESS_DELETE	C $\leftarrow$ S
5	ZEBRA_INTERFACE_UP	C $\leftrightarrow$ S
6	ZEBRA_INTERFACE_DOWN	C $\leftrightarrow$ S
7	ZEBRA_IPV4_ROUTE_ADD	C $\leftrightarrow$ S
8	ZEBRA_IPV4_ROUTE_DELETE	C $\leftrightarrow$ S
9	ZEBRA_IPV6_ROUTE_ADD	C $\leftrightarrow$ S
10	ZEBRA_IPV6_ROUTE_DELETE	C $\leftrightarrow$ S
11	ZEBRA_REDISTRIBUTE_ADD	C $\rightarrow$ S
12	ZEBRA_REDISTRIBUTE_DELETE	C $\rightarrow$ S
13	ZEBRA_REDISTRIBUTE_DEFAULT_ADD	C $\rightarrow$ S
14	ZEBRA_REDISTRIBUTE_DEFAULT_DELETE	C $\rightarrow$ S
15	ZEBRA_IPV4_NEXTHOP_LOOKUP	C $\rightarrow$ S
16	ZEBRA_IPV6_NEXTHOP_LOOKUP	C $\rightarrow$ S
17	ZEBRA_IPV4_IMPORT_LOOKUP	C $\rightarrow$ S
18	ZEBRA_IPV6_IMPORT_LOOKUP	N/A
19	ZEBRA_INTERFACE_RENAME	N/A
20	ZEBRA_ROUTER_ID_ADD	C $\rightarrow$ S
21	ZEBRA_ROUTER_ID_DELETE	C $\rightarrow$ S
22	ZEBRA_ROUTER_ID_UPDATE	C $\leftarrow$ S

of the protocol. The value of this field should equal 255. In the older version of the protocol this field was interpreted as a command. The following 8-bit field **version** specifies the version of the Zebra protocol. The present version of the protocol is 1. The last field of the header field is a 16-bit field **command** that specifies the command. Command code is written in network order. Content of the field **message data** depends on the command. In the Zebra protocol there are 22 messages provided, listed in Table I. The first column specifies the value of the message code that is placed in the field **command**. The second column contains the name of the message, and the third one the direction in which it is sent. The letter C stands for a process running in the application layer, and the letter S means a module running in the layer within the Zebra equipment. These messages can be sent to the Zebra module (direction C $\rightarrow$ S), to a module with routing process (direction C $\leftarrow$ S), or in both directions (C $\leftrightarrow$ S). Messages sent in both directions often have asymmetric forms. Therefore, it is necessary to use the right tool for an analysis of sent messages. In this case, it may be the Wireshark program. Unfortunately, the implemented module in the above mentioned program is for the analysis of the older, now outdated, version of the protocol. Therefore, a Wireshark modification, that includes the new version of the Zebra protocol, needs to be prepared for the construction of the test platform.

The process supporting the routing connects to the Zebra module. Zebra server, running in the Zebra module, may support many connections simultaneously. It should be noted that not all messages included in the Zebra protocol have been implemented. Those not implemented are marked in the column specifying the direction as N/A.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

The module supporting the routing protocol is running on an abstract hardware and, in this way, it is partially independent from the operating system. The module is not fully independent because in order to send or receive a signal, such as Ospf Hello message for OSPFv3 protocol, direct use of the socket API is required. By sending a signal message via the API operating system, the routing protocol specifies an interface through which the message is to be sent. Similarly, it also uses the API operating system to read messages. The system returns the received message, the source and destination address, as well as the information on the interface that received the message. This requires an application of a particular function from the API operating system. Some differences in functioning of the socket API may occur, depending on the operating system. This forces the adjustment of a program, running at the application layer, to a specific operating system.

The architecture of the Quagga project, optimized with regards to its performance, has some disadvantages: no process can be moved on another machine and the implementation of routing protocols is dependent on the API operating system. In order to make the mentioned platform applicable for the analysis and testing of existing routing protocols and to design new routing protocols (also for teaching purposes), a modification that enables a physical division of the Data Plane (DP) and the Control Plane (CP) functions is proposed in the next section.

III. A CONCEPT OF A PLATFORM FOR TESTING AND DEVELOPING ROUTING ALGORITHMS

The main idea of the proposed platform is to move the application layer to a dedicated computer. In the proposed solution, the application layer meets the CP functionality, and the hardware layer implements the functionality of the DP. A similar approach was applied to the GMPLS system [5], introducing a distinction between CP and DP. In comparison to the GMLS system, the difference is that CP in the proposed solution does not have its own signaling network. The routing protocol messages (supported by CP) are sent via network, supported by DP. Owing to this division, the application layer does not use the hardware resources directly. In this way, one machine (virtual or physical) can support many application layers concurrently, each for a separate node.

Figure 5 presents a network consisting of N nodes. Their CP is moved to a separate machine, shared by all nodes. Each router shown in Figure 5 has a dedicated router implementing the DP functions, while the machine supporting CP has many CP instatnces running. Each such instance is shown as a rectangle drawn with a dotted line. The proposed network may have one central computer that is running numerous CP instances for all nodes (as shown in Figure 5), or CP can be distributed across different machines.

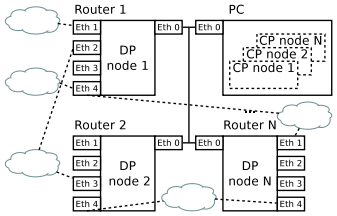


Figure 5. The proposed platform for testing and developing routing algorithms

In the extreme case, each node has a separate computer implementing the CP functions[1].

In order to connect CP with DP, a separate network was dedicated for this purpose. In Figure 5 all devices implementing DP functions are connected to this network via eth0 interface. This interface is unavailable for the DP network and invisible for the routing protocols. The routing protocols specify the path for the DP networks that consist of the nodes, using the eth1–eth4 interfaces. The computer with CP instances is connected to the network that supports an interface between CP and DP (in Figure 5 via eth0 interface).

Each CP instance must be properly configured. The configuration specifies the eth0 interface address of the DP device that is supported by a corresponding instance of CP. A single CP instance is composed of multiple processes, each supporting a different routing protocol. The routing protocol can be configured using the CLI. Access to the console is controlled using the telnet protocol. Any process that supports routing process listens for TCP connections on the specified port. This port must be configured so as to be unique within a given machine. This requires additional settings.

The Quagga software router processes running save their IDs in file /var/run/quagga. The file name depends on the routing protocol, supported by a given process. It allows for a simple stops/starts of the processes. Supporting of multiple processes for the same routing protocol requires a unique name of the file stored in the folder /var/run/quagga for each of the processes. A unique file name should depend on the routing protocol and the node on which the protocol operates. The last parameter that must be set is the name of the file where the configuration of routing protocol, designed for a given node, is stored.

In summary, the computer supporting multiple CP instances have multiple processes with the same name running, which are serving the same routing protocols. For each process, the following items have to be set:

- IP address of DP device,
- Unique port number at which a telnet server is listening,
- Unique filename of file, where process ID is stored,
- Unique filename of file that stores configuration of

[1]An idea of a central CP was proposed in an OpenFlow project [6]

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

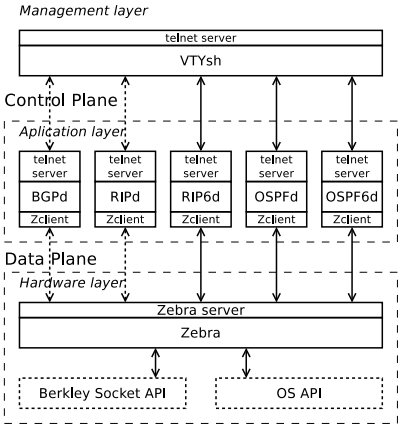


Figure 6. The node architecture in the proposed platform

routing protocol.

The presented approach has many advantages, despite the need to provide additional parameters for each of the processes. The entire network configuration is stored on one machine. All consoles to configure the routing protocols are available from the same machine, so there is no need for the management layer. There is also no need to output the console cables for all nodes. The modification of the process supporting the routing protocol is easier -– it only needs to be compiled once and then restart a modified process for all nodes. All of these steps can be implemented with access to a single computer.

The processes take the mentioned parameters from the arguments with which the process was started. This allows to write a simple script that starts, restarts, or turns off the processes for a single or all nodes. The following is a sample script that runs routing protocol OSPFv3 for node 1 with an IP address eth0 10.0.1.1.

```
./ospf6d --demonize \
 --dpaddr 10.0.1.1 --cliport 2701 \
 --pid /var/run/quagga/n1_ospf6d.pid \
 --conf /etc/quagga/n1_ospf6d.conf
```

To stop the OSPFv3 process for node 1, the following script needs to be started.

```
kill 'cat /var/run/n1_ospf6d.pid'
```

Figure 6 presents the node architecture (software router) in the proposed platform for testing and developing the routing protocols. The top rectangle drawn with a dotted line includes the processes responsible for the CP functions, and the bottom one includes the processes responsible for the DP. Each CP process communicates via the Zebra protocol with a Zebra module. The Zebra module supports API that controls the work of DP. The Zebra server is responsible for communication with a higher layer. In the proposed node, the modification the Zebra module additionally supports sending and receiving of signaling messages of routing protocol. The implementation of these steps required a Zebra protocol to be modified. For this purpose, two new messages were added. They are presented in Table II. Sending and receiving of a Zebra protocol message follows via ZEBRA_TRANSMIT command. Figure 7 presents a format of such message. A message length depends on the length of a packet (IPv4 / IPv6) with the signaling message that we want to send. The first parameter **interface idx** in the data field is a 32-bit value with an index of the interface through which a message is to be sent. The index of the interface is the same as the one in the message, adding a new interface. Interface index was saved as 32-bit value intentionally, since all the interface indexes in API of the Linux system [7] are written in the form of 32-bit numbers. The final element of the message is a IPv4 or IPv6 packet (including the header). In the header of the transmitted packet a destination IP address is stored. In the case of IPv6, the packet being sent should include the calculated checksum before sending it, although it is possible that DP counts the sum. Receiving a signaling messages from the network by CP requires DP mediation. DP needs to know the destination address of the packet that CP wants to receive. To do so, the right filter must be set, where unicast or multicast address is specified. The command ZEBRA_CONFIGURE_RECEIVER helps to configure DP. After an appropriate configuration, the DP transmits the received messages to CP via ZEBRA_TRANSMIT command. The format of this message is symmetric and remains the same regardless of the direction.

Table II
THE NEW MESSAGES OF ZEBRA PROTOCOL

code	command	dir
23	ZEBRA_CONFIGURE_RECEIVER	C –> S
24	ZEBRA_TRANSMIT	C <–> S

+	Bits 0-7	Bits 8-15	Bits 16-23	Bits 24-31
0	Message size		255	1
32	24		Interface Idx	
64	Interface Idx			
96	IPv4 / IPv6 packet with header			
...				

Figure 7. The format of ZEBRA_TRANSMIT message

IV. IMPLEMENTATION

The implementation is in progress. The aim of the DP implementation is to build a firmware image which is then flashed into the router. Linux is frequently chosen as an embedded system. The firmware image of embedded Linux consists of a kernel and a file system. Open Embedded [8] is a toolset and sources for embedding Linux. The set offers many possibilities, ranging from a kernel with a basic toolkit up to a system with a graphic interface. With regards to the network devices, Open WRT [9] distributions have been developed. It includes a toolset for an oblique compilation,

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

the addresses of repositories with a system and programs kernel sources, and a set of patches that allow kernel or programs adjustment to the given hardware platform. The software set has been narrowed. An Open WRT is simpler in implementation, as compared to Open Embedded, and a software developer with a little experience can easily build a firmware image, using a creator. The configurator attached to the project can be run, using the `menu config` command. It enables a choice of hardware platform, a device and a set of programs. Moreover, it makes an addition of one's own programs possible. In order to build a firmware image for DP, the code with a Zebra module from Quagga project needs to be added and then modified. The modifications consist in adding the commands and functions to send and receive the messages or routing protocols described in the previous section.

The implementation of CP software consists in downloading of a Quagga project source code and its further modifications. The modification is based on adding the new commands to the Zebra protocol. It is necessary to modify the way of sending and receiving of messages for each routing algorithm. These operations are executed via the Zebra module. Thus, in order to send the packet via a given interface, its number and packet need to be placed in a message data field compatible with Zebra protocol. Similarly, packet reception is activated after Zebra module message is received. This message includes in its data field an interface number and the packet that has been received by this interface. The functions for sending and receiving packets are stored in file x_network.c, where x stands for a process name, e.g., ospf6d for the OSPFv3 protocol.

It needs to be noted that the system will run in the same way as the Quagga software router, if both softwares, DP (Zebra module) and CP (the modules with routing algorithms) are installed on the same machine. That means that the proposed code modification, after being implemented, may be added to the Quagga project. As a result, it will be possible to disperse one node onto more machines, to place all the processes supporting the routing protocol on one machine and to port very easily a project onto the new operating systems (all that needs to be done is to modify the Zebra module).

V. USAGE SCENARIOUS

The proposed platform enables fulfilling the labs with routing protocols like RIP or OSPF. Regardless of the routing algorithm, the labs include the following tasks:

- Preparing physical connections between DP devices;
- Connecting the CP device to the network, and checking the communication between CP and DP devices;
- Launching the routing protocol, e.g. RIP;
- Watching the entries in forward (routing) table;
- Checking if the network is working correctly (ping command);
- The analysis of exchanged routing protocols' messages;
- The analysis of messages exchanged between CP and DP;
- Checking if the network is able to establish a new path afer physical breaking the link (disconnecting the cable);
- Making the changes in routing protocols configuration, e.g. changing the timers in the RIP protocol.

VI. CONCLUSION

This article presents a new concept of a network platform that enables both an analysis of existing network protocols and the implementation, plus testing, of new protocols. The proposed platform, based on the Quagga software router concept, combines the advantages of solutions relying on the network node vitualization with the testing networks implemented with the help of hardware routers. The platform enables a simple addition and modification of routing protocols in a testing network without a need for cross-compilation and uploading the firmware to each of the routers. Simultaneously, the complexity of a routing protocol does not affect the functioning of a hardware node. This allows to implement the platform using much simpler and cheaper hardware nodes. In further works, related to the proposed platform for testing the routing protocols, an analysis module for the Wireshark program will be developed. This should allow to pick up an information exchange with hardware nodes on the CP interface.

REFERENCES

[1] "Quagga homepage." [Online]. Available: www.quagga.net/ <retrieved: May, 2012>

[2] A. Bianco, R. Birke, J. Finochietto, L. Giraudo, F. Marenco, M. Mellia, A. Khan, D. Manjunath, "Control and management plane in a multi-stage software router architecture," in *High Performance Switching and Routing*, May 2008, pp. 235–240.

[3] "Xorp homepage." [Online]. Available: www.xorp.org/ <retrieved: May, 2012>

[4] "Xorp architecture." [Online]. Available: http://xorp.run.montefiore.ulg.ac.be/latex2wiki/design_overview <retrieved: May, 2012>

[5] E. Mannie, "Generalized Multi-Protocol Label Switching (GMPLS) Architecture," RFC 3945 (Proposed Standard), Internet Engineering Task Force, Oct. 2004.

[6] "Openflow switch specyfication," Feb. 2011. [Online]. Available: http://www.openflow.org/documents/openflow-spec-v1.1.0.pdf <retrieved: May, 2012>

[7] *The Linux Kernel*. [Online]. Available: http://kernelbook.sourceforge.net/ <retrieved: May, 2012>

[8] "OpenEmbedded." [Online]. Available: http://www.openembedded.org <retrieved: May, 2012>

[9] "OpenWRT." [Online]. Available: https://openwrt.org/ <retrieved: May, 2012>

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Telephony Fraud Detection in Next Generation Networks

Simon Augustin, Carmen Gaißer, Julian Knauer, Michael Massoth, Katrin Piejko, David Rihm and Torsten Wiens
Department of Computer Science
Hochschule Darmstadt
Darmstadt, Germany
senily64dx@googlemail.com, carmen.gaisser@stud.h-da.de, jpk@goatpr0n.de, katrin.piejko@stud.h-da.de,
david.rihm1@freenet.de, michael.massoth@h-da.de, torsten.wiens@h-da.de

***Abstract*—Telephony fraud is a growing problem for telecommunication service providers that operate Next Generation Networks (NGN). This paper describes a framework for a rule-based fraud detection system. The classification of fraudulent calls is based on Call Detail Records (CDR) that are used by telecommunication service providers for billing purposes. By analyzing this data, fraud can be detected efficiently. We propose a method for accomplishing this. The work has been conducted in collaboration with a telephony service provider that made real-life CDR data available for analysis. The main achievement of this paper is the description of a rule-based system that detects telephony fraud using CDR data.**

Keywords-Communication system security; Communication system signaling; Communication system traffic; Computer network management; Next generation networking

I. INTRODUCTION

Telephony fraud is a serious problem for carriers that operate Next Generation Networks (NGN). Attackers regularly try to compromise accounts of users or providers to circumvent charging systems or to cause financial harm to customers. Telephony fraud comprises unauthorized deletion or alteration of billing records, unauthorized bypassing of lawful billing systems, unauthorized billing and the taking of service provider property [1].

A. Current situation

The Communications Fraud Control Association (CFCA) estimated in 2009 that fraud leads to a worldwide annual loss of 74 to 80 billion USD [2]. It is expected that this value will increase in the future. The top three fraud types, as named in their report, are (see Figure 1):

- Subscription or identity theft (22.0 billion USD)
- Compromised Private Branch Exchange (PBX) systems (15.0 billion USD)
- Premium rate service fraud (4.5 billion USD)

Even single fraud attacks may cause significant losses. In one case, an attacker conducted 11,000 calls to Australia, causing an estimated damage of more than 120,000 USD. These calls were made over a period of only 46 hours [3]. These losses could be drastically reduced if effective real-time fraud detection mechanisms were applied.

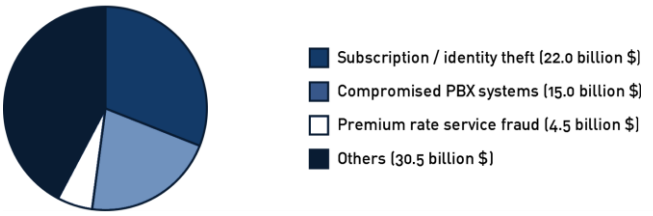


Figure 1. Top three fraud types

This kind of fraud also causes significant economic damage because some small- and medium-sized enterprises (SME) may not be able to deal with the amount of financial damage caused by these attackers, possibly leading to bankruptcy.

B. Challenges in fraud detection

In order to develop well performing fraud detection mechanisms, access to real world data is necessary. However, telecommunication providers are not allowed to expose this data due to privacy reasons. This is caused by national legal limitations, for example the German "Bundesdatenschutzgesetz" (Federal Data Protection Act) [4]. Additionally, fraud detection is not just a binary problem. The precise classification of calls as fraudulent or not with a minimum of false positives is difficult. There are cases that cannot be decided with certainty. Therefore, fraud detection has to be treated as an n-class problem [5].

C. Structure of the paper

This paper is structured as follows: Section II gives an overview on the recent activities in the field of fraud detection. Section III describes the basic concept of fraud detection and our design decisions for the framework. After the fundamentals have been explained, a more detailed description of our approach is given in Section IV. The paper ends with a conclusion and an outlook on future work in Section V. Acknowledgements follow in the last section.

II. RELATED WORK

In this paper, a rule-based system for fraud detection is described. The field of fraud detection can be divided into multiple categories. Two important ones are rule-based approaches and neural networks. There are also additional approaches, for example Bayesian Networks, Support Vector

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

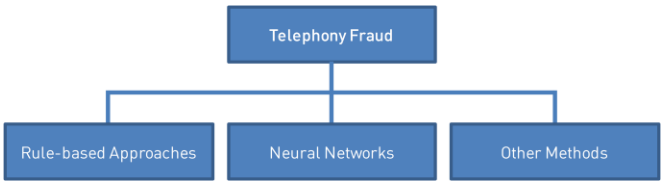


Figure 2. Methods to counter telephony fraud

Machines and Hidden Markov Models. These are described in Section II.C (see Figure 2).

A. *Rule-based methods*

Rule-based methods are very effective, but hard to manage. Extensive work is required to specify rules for every imaginable fraud case. Another downside is that rule-based fraud detection systems need to be updated frequently to cover new kinds of fraud [6].

Rosset et al. [7] proposed an extension of the C4.5 algorithm that divides a rule-discovery process into two steps. The first step generates a large number of candidate rules. The second step puts together a rule-set from these candidates. Olszewski [8] constructed a detection method based on user profiling by employing the Latent Dirichlet Allocation (LDA). Using the Kullback-Leibler divergence, the participants are classified as "good" or "evil". Ruiz-Agundez et al. [9] propose an architecture for rule-based mechanisms that can be applied on NGN infrastructures.

B. *Neural networks*

One of the alternatives to rule-based approaches for classification are neural networks. These are more suitable to cover new and unknown attacks. Taniguchi et al. [10] summarize three methods for fraud detection, one being a neural network. They claim that these three types are able to detect 85% of all fraud cases that occurred in their test set.

1. The first method consists of the application of a feed-forward neural network. It is used to learn a discriminative function to classify service subscribers using summary statistics.
2. The second method applies a Gaussian mixture model to determine the probability of the user's future behavior. This is based on user behavior in the past. The probabilities are used to validate the current behavior in order to detect deviations.
3. The third method uses a Bayesian network. Here, statistical properties of users and of multiple fraud cases are used.

The application of neural networks for fraud detection in mobile communication has been introduced by Qayyum et al. [11]. A disadvantage of their approach is that further adjustments are needed for the system in order to work efficiently.

C. *Other methods*

The pattern recognition skills of the human eye are very powerful. Therefore, Cox et al. [12] proposed to apply humans in the process of fraud detection. They introduced multiple techniques to visualize network traffic in a human readable way. Hollmén and Tresp [13] proposed a system that is based on a hierarchical regime-switching model. This system receives inference rules from a junction tree algorithm and is trained by using the Expectation Maximization (EM) algorithm.

III. CONCEPT AND OVERALL SYSTEM DESIGN

Every internet telecommunication service provider uses charging systems that log each call that was made using the network of the service provider. These log files contain detailed information about calls, and are commonly referred to as Call Detail Records, or sometimes as Call Data Records (CDR). In the CDR, the subscriber numbers of caller and callee, the date and time when the call was made and the call duration are recorded. Therefore, these log files contain valuable information that can be used to detect telephony fraud. Since CDR data is not allowed to be exposed to the public because of German legal regulations, the data provided by the cooperating telecommunication service provider had to be anonymized.

Our system uses CDR files and analyzes them for anomalies (see Figure 3). This is accomplished by different filters. Each filter scans the CDRs using specific rules. If an anomaly is detected, and one of the filters supplies a positive result, there is a strong suspicion that a fraud case has occurred. This fraud case has to be validated by a human and further actions, for example the temporary deactivation of an account, have to be taken. Our framework does not automatically perform these actions, as telephony fraud comprises false positives.

The framework has been implemented in Python 2.7. The decision to use Python resulted from several considerations. First of all, Python can be learned quickly and, due to its code structure, is easy to read. This ensures a quick start of implementation and results in low costs for later maintenance and the addition of extensions. Furthermore, Python is an open source product that is highly portable and runs on almost every operating system [14].

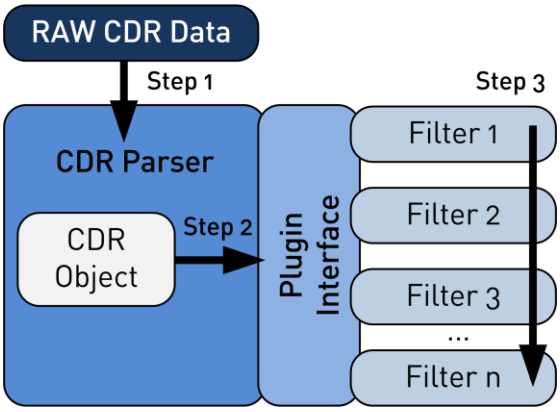


Figure 3. System overview

IV. SYSTEM COMPONENTS

In this section, the system components are described in detail.

A. Structure of a CDR

Each CDR consists of several elements that correspond to different functionalities. These elements indicate the start and the end of a call, among other parameters. Each element contains the date and time when the element was written. The first element, indicating the beginning of a record, contains the unique session ID that identifies a CDR. The elements that are necessary for further analysis are now described in more detail.

The Incoming element of a CDR (called A-element in the CDR specification) contains the properties of an incoming call [15]. For our purpose, only the carrier ID (n-attribute of the A-element) is important.

The Connected element (C-element) only exists if a conversation was established. The C-element consists of several sub-elements. For example, its x-element contains the Session Initiation Protocol-(SIP) data of the connection. The SIP data contains several fields, starting at position zero. The first field corresponds to the number of the callee. The 13th and 25th field both contain the customer ID or the subscriber number. Furthermore, the C-element includes the duration of a call in milliseconds.

If a call lasts longer than 15 minutes, the CDR is split into multiple parts. These parts can be identified by the first number in the S-element. This element is the first element in a CDR, indicating the beginning of the CDR. If the call duration is below 15 minutes, the identifier is set to ”0”. If it indicates the start of a record series, it is set to ”1”. The final part is marked ”3”. All parts in between are set to ”2”.

If a call is finished, the Disconnecting element (D-element) is written. In this element, the reason for the call’s termination is stored. The From-field in this element is also important, as it indicates which party hung up. In a nutshell, the C- and the D-element provide the necessary information to bill a call.

B. Framework

To analyze the CDRs, we developed a framework that is capable of parsing the log files generated by the billing system. The framework consists of multiple parts:

- Classes for CDRs and CDR-elements into which the input data is parsed.
- The main part of the software that controls the application flow.
- Several filters implementing the rules for fraud detection.

Now, the individual parts of the framework are explained in more detail.

1. CDR Classes: The framework contains classes for each CDR element (see previous section). This modular structure provides easy filter access to the different CDR elements.

2. Main part: This part of the software controls the application flow. It starts the application, evaluates the console commands for the input files that are to be parsed and registers the different filters. The filters are organized as a list, which is iterated for each input CDR. To expand the software, more filters can easily be integrated into the analysis process, simply by adding them to the list of registered filters.

The CDR parser starts to read the data from the given input files. Each CDR is parsed from the log files into a CDR object. Each filter expects a CDR object as input and analyzes it. After the input files have been parsed completely, the results from the filters are collected by the main part. If one filter or multiple filters have detected a potential fraud case, the output is saved to a text file. In this case, an operator is alarmed.

The release candidate comes as a console application. A graphical user interface has not been included, since the software is used by the technical staff of the cooperating telecommunication service provider and the systems that process the CDRs are UNIX-based. Hence, a command line interface is sufficient.

C. Filters

The framework includes a filter base class that is inherited by all implemented filters (see Figure 4). This base class contains methods for all filters, e.g., for the formatting of date and time, and a method that returns the results. For each rule, which was defined to detect fraud, a filter is implemented. Each filter analyzes a given CDR, evaluates it for fraud-suspicious data and returns the collected results to the main class.

In general, all filters only regard calls originating from the internet telecommunication service provider's network, as only these calls are charged. These are identified if the callee’s subscriber number corresponds to a customer ID and the carrier ID in the Incoming element of the CDR does not correspond to the service provider’s ID.

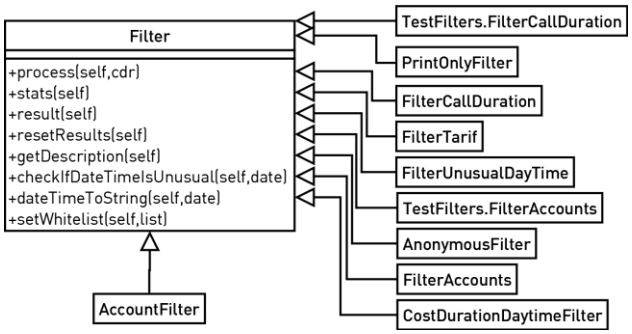


Figure 4. Filter base class and inherited classes

Up to date, four filters have been developed. The first filter regards only single calls of a customer. The second one regards all calls of a specific customer per hour. The third filter scans for signaling errors and suppressed caller IDs, while filter number four considers historical user data.

The first filter analyzes a single call for the following criteria:

- The duration of the call, depending on the destination pay scale area.
- The date and time when the call was made.

To classify the pay scale area, the destination area code of the callee's subscriber number is analyzed. We defined four categories of pay scale areas:

1. No charges: The first category classifies calls that only cause low charges or none at all. Therefore, these calls are omitted. As the software was developed in cooperation with a German company, the relevant area codes include the German fixed network, Voice over Internet Protocol (VoIP) and national subscriber numbers.
2. Moderately expensive: This category comprises calls destined for the German mobile network. These calls are not very expensive, regarding the charges per minute. In this case, calls lasting for more than a specific threshold are considered unusual.
3. Expensive: To simplify the classification, this category includes all calls that do not belong to one of the other categories. These are calls that are destined for international and special rate numbers. A threshold for the call duration is set accordingly.
4. Very expensive: Satellite calls belong to the most expensive category. These calls may be charged at up to 20 € per minute. Therefore, the threshold in this category is considerably lower than the thresholds in the previous categories. The second criteria for this filter are the date and the time when the call took place. If, for example, a company only has business customers, it can be assumed that calls outside the business hours or on weekends are more suspicious than others.

The second filter regards all calls that are made by a specific customer in a given time frame. The criteria are as follows: If the amount of calls per hour is greater than a specific value or if the overall call duration per hour exceeds a specific threshold, it is assumed that this is a fraudulent usage of the telephony service.

The first and the second filter also include a whitelist for specific customers. Whitelist candidates are customers who would regularly be above the thresholds with their normal call behavior, and therefore would be considered as fraudulent. Those customers are maintained in the whitelist and are ignored by the filters.

The third filter scans the input data for signaling errors and suppressed caller IDs, since these may also denote fraud cases. These parameters are only considered for analysis if they are found on incoming calls. Additionally, data in the CDRs indicating the connection quality is assessed by this filter. One of the typical fraud scenarios consists of routing calls via multiple international service providers. In these cases, connection quality may drop significantly. Therefore, low connection quality may be another indicator for fraud cases.

The fourth filter collects historical user data, for example the total duration of calls made by a single user or by all users. Here, up to seven categories may be included. Additionally, this filter is able to output descriptive statistics and diagrams as a PDF file.

Another interesting information in a CDR is the reason for call termination, which is stored in the D-element. Among the possible reasons, SIP and identity errors are the most interesting ones from the perspective of fraud detection. These reasons can also be used for statistical purposes or to detect internal network errors.

The filter rules and their associated thresholds have been determined by a thorough evaluation of actual fraud cases. This has been actively supported by the collaborating service provider. Unfortunately, it is not possible to describe the rules and thresholds in more detail. A publication of these parameters would give attackers a significant advantage in bypassing the system, which is productively used.

D. Conclusion and future work

In general, the presented rule-based approach for detecting telephony fraud is promising. The described solution performs well on the real-life CDRs delivered by the service provider, regularly classifying about 4% as false positive fraud cases. Additionally, it is almost an order of magnitude faster than the solution previously used, which was script-based. For example, the presented system is able to process typical CDR files in significantly less than one minute, while the old system took more than ten minutes to accomplish this, under identical circumstances. Furthermore, the system did not only detect known fraud attacks, but also discovered yet unknown signaling errors that were caused by other carriers. Future work will comprise an investigation of these signaling errors, since they appear to be potential predictors for telephony fraud. This especially concerns so-called inter-carrier fraud.

Still, the developed system needs more testing. It appears that the thresholds have to be specified more precisely. As these values rely on experiences, the software has to be run in a productive environment with near real-time data to exactly determine the thresholds, in order to increase the detection probability. The final decision, if the results detected by the system are fraud, still relies on a human operator judging each case. Much harm could be done by automatically blocking innocent customers due to false positive classification results. With the presented approach, our system is able to conduct most of the analysis necessary to detect fraud by itself. Therefore, the probability that the delivered results indicate real fraud cases is already high.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1

Given the modular implementation, the system can be easily extended. More rules, that is to say more filters, can be integrated with no effort. The more distinct the filters are that analyze the incoming data, the more likely it is to detect fraud before too much damage is done.

Granted that the presented system is tested more thoroughly, it will be capable to be used on a Next Generation Network for performant fraud detection. Its application will possibly improve the detection of telephony fraud, and it is worth considering for use by telecommunication service providers. From the collaborating service provider's perspective, the presented approach represents a major achievement concerning fraud detection in their practice, compared to the previously used solution.

ACKNOWLEDGMENT

This work has been performed for the “Fraud Detection” project of Hochschule Darmstadt - University of Applied Sciences. The project is partially funded by the “Bundesministerium für Bildung und Forschung” (BMBF) and supported by ”Center for Advanced Security Research Darmstadt” (CASED). The authors additionally would like to acknowledge the support of toplink GmbH, which made this work possible.

REFERENCES

[1] Zar, J. et al., "VOIPSA - VoIP security and privacy threat taxonomy, public release 1.0", http://www.voipsa.org/activities/taxonomy.php, October 2005.

[2] Communications Fraud Control Association, “2009 global fraud loss survey,” http://www.cfca.org/, 01. 09. 2011.

[3] S. Tindal, “VoIP hackers strike perth business,” ZDNet, Jan. 2009. http://www.zdnet.com.au/voip-hackers-strikeperth-business-339294515.htm, 05. 08. 2011.

[4] Bundesministerium der Justiz, „Bundesdatenschutzgesetz in der Fassung vom 14. Januar 2003, zuletzt geändert am 14. August 2009,“ Berlin, 2009.

[5] T. Padmaja, N. Dhulipalla, R. S. Bapi, and P. R. Krishna, “Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection,” in: Proceedings of the 15th International Conference on Advanced Computing and Communications (ADCOM 2007). IEEE Computer Society, 2007; pp. 511–516.

[6] Y. Kou, C.-T. Lu, S. Sirwongwattana and Y.-P. Huang, “Survey of fraud detection techniques,” in: Proceedings of the 2004 IEEE International Conference on Networking, Sensing and Control (ICNSC 2004). IEEE, 2004; pp. 749–754.

[7] S. Rosset, U. Murad, E. Neumann, Y. Idan and G. Pinkas, “Discovery of fraud rules for telecommunications challenges and solutions,” in: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 1999). ACM, 1999; pp. 409–413.

[8] D. Olszewski, “Fraud detection in telecommunications using Kullback-Leibler divergence and latent Dirichlet allocation,” in: Proceedings of the 10th International Conference on Adaptive and Natural Computing Algorithms (ICANNGA 2011). Springer, 2011; pp. 71–80.

[9] I. Ruiz-Agundez, Y. Penya and P. Garcia Bringas, “Fraud detection for voice over ip services on next-generation networks,” in: Proceedings of the 4th Workshop in Information Security Theory and Practice (WISTP 2010). Springer, 2010; pp. 199–212.

[10] M. Taniguchi, M. Haft, J. Hollmén 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 (ICASSP 1998). IEEE, 1998; pp. 1241–1244.

[11] S. Qayyum, S. Mansoor, A. Khalid, K. Khushbakht, Z. Halim and A. Baig, “Fraudulent call detection for mobile networks,” in: Proceedings of the 2010 International Conference on Information and Emerging Technologies (ICIET 2010). IEEE, 2010.

[12] K. C. Cox, S. G. Eick, G. J.Wills and R. J. Brachman, “Visual data mining: Recognizing telephone calling fraud,” in: Data Mining and Knowledge Discovery, vol. 1, no. 2, pp. 225–231, Jun. 1997.

[13] J. Hollmén and V. Tresp, “Call-based fraud detection in mobile communication networks using a hierarchical regime-switching model,” in: Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems 11 (NIPS 1999). Morgan Kaufmann, 1999; pp. 889–895.

[14] P. S. Foundation, “Python programming language - official website,” http://www.python.org, 1990-2011.

[15] TELES, “Teles.icdr, S48-S2000 series,” Teles Communication Systems, 2006.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-199-1