AICT 2015

The Eleventh Advanced International Conference on Telecommunications

ISBN: 978-1-61208-411-4

June 21 - 26, 2015

Brussels, Belgium

AICT 2015 Editors

Eugen Borcoci, University Politehncia Bucharest, Romania

Tulin Atmaca, Telecom SudParis, France

AICT 2015

Foreword

The Eleventh Advanced International Conference on Telecommunications (AICT 2015), held between June 21-26, 2015, in Brussels, Belgium, covered a variety of challenging telecommunication topics ranging from background fields like signals, traffic, coding, communication basics up to large communication systems and networks, fixed, mobile and integrated, etc. Applications, services, system and network management issues also received significant attention.

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

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

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

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

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

We take here the opportunity to warmly thank all the members of the AICT 2015 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to AICT 2015. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the AICT 2015 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that AICT 2015 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of telecommunications.

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

AICT 2015 Chairs:

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

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

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

AICT 2015

COMMITTEE

AICT Advisory Committee

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

AICT Industry/Research Chairs

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

AICT Publicity Chair

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

AICT 2015 Technical Program Committee

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

Andres Arjona, Nokia, Japan
Michael Atighetchi, Raytheon BBN Technologies-Cambridge, USA
Tulin Atmaca, TELECOM SudParis, France
Konstantin Avratchenkov, INRIA- Sophia Antipolis, France
Hajer Bargaoui, University of Burgundy, 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
Robert Bestak, Czech Technical University in Prague, Czech Republic
Antonella Bogoni, CNIT (Inter-University National Consortium for Telecommunications), Italy
Eugen Borcoci, University "Politehnica" of Bucharest (UPB), Romania
Christos Bouras, University of Patras, Greece
Salah Bourennane, Ecole Centrale Marseille - Institut Fresnel, France
Lubomir Brancik, Brno University of Technology, Czech Republic
Peter Brida, University of Zilina, Slovakia
Julien Broisin, Université Paul Sabatier, Toulouse III, France
Damian Bulira, Wroclaw University of Technology, Poland
Prasad Calyam, The Ohio State University, USA
Maria-Dolores Cano Banos, Universidad Politécnica de Cartagena, Spain
Daniel Carvalho da Cunha, Federal University of Pernambuco - UFPE, Brazil
Fernando Cerdan, Universidad Politecnica de Cartagena, Spain
Júlio Cesar Nievola, Pontificia Universidade Catolica do Parana (PUCPR), Brazil
Hakima Chaouchi, Telecom SudParis, France
Amitava Chatterjee, Jadavpur University, India
Phool Singh Chauhan, Indian Institute of Technology Kanpur, India
Rajesh Chharia, CJ Online PVT. LTD., India
Stefano Chessa, University of Pisa, Italy
Carlo Ciulla, University for Information Science and Technology, Republic of Macedonia
Sungsoo Choi, Korea Electrotechnology Research Institute (KERI), S. Korea
Richard G. Clegg, University College London, UK
Johanne Cohen, LRI, France
Todor Cooklev, Indiana-Purdue University - Fort Wayne, USA
Dimitrios Damopoulos, Stevens Institute of Technology, USA
Arnaud de La Fortelle, MINES ParisTech, France
Flávio de Oliveira Silva, Federal University of Uberlandia (UFU), Brazil
Chérif Diallo, Consultant Sécurité des Systèmes d'Information, France
Fábio Diniz Rossi, Farroupilha Federal Institute of Education, Science and Technology, Brazil
Sourav Dutta, Max-Planck Institute for Informatics, Germany
Zbigniew Dziong, École de Technologie Supérieure - Montreal, Canada
Ghais El Zein, IETR - INSA Rennes, France
Mohamed El-Tarhuni, American University of Sharjah , UAE
Anna Esposito, Second University of Naples, Italy
Mário Ezequiel Augusto, Santa Catarina State University, Brazil
Mário F. S. Ferreira, University of Aveiro, Portugal
Bruno Filipe Marques, Polytechnic Institute of Viseu, Portugal
Claudio Forlivesi, Bell Labs - IoT, Belgium

Pedro Fortuna, University of Porto, Porutgal
Paraskevi Fragopoulou, TEI of Crete, Greece
François Gagnon, Cégep de Sainte-Foy, Canada
Alex Galis, University College London, UK
Seema Garg, Hughes Systique Corporation, USA
Rung-Hung Gau, National Chiao Tung University, Taiwan
Bezalel Gavish, Southern Methodist University Dallas, USA
Christos K. Georgiadis, University of Macedonia - Thessaloniki, Greece
Mircea Giurgiu, Technical University of Cluj-Napoca, Romania
Mariusz Glabowski, Poznan University of Technology, Poland
Savo Glisic, University of Oulu, Finland
Claire Goursaud, INSA de Lyon - Villeurbanne, France
Roger J. Green, University of Warwick, UK
Stefanos Gritzalis, University of the Aegean, Greece
Vic Grout, Glyndwr University - Wrexham, UK
Carlos Guerrero, University of the Balearic Islands, Spain
Lei Guo, Northeastern University, China
Jan Haase, University of the Federal Armed Forces Hamburg, Germany
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
Ilias Iliadis, IBM Zurich Research Laboratory, Switzerland
Muhammad Ali Imran, University of Surrey - Guildford, UK
Lucian Ioan, University: "Politehnica" of Bucharest (UPB), Romania
Henric Johnson, Blekinge Institute of Technology, Sweden
Peter Jung, University Duisburg, Germany
Michail Kalogiannakis, University of Crete, Greece
Georgios Kambourakis, University of the Aegean - Samos, Greece
Dimitris Kanellopoulos, University of Patras, Greece
Charalampos Karagiannidis, University of Thessaly - Volos, Greece
Ziad Khalaf, SUPELEC/SCEE, France
Mehdi Khouja, University of Gabes, Tunisia
Kashif Kifayat, Liverpool John Moores University, UK
Insoo Koo, University of Ulsan, Korea
Francine Krief, LaBRI, France
Robert Koch, University of the Federal Armed Forces / German Navy, Germany
Dragana Krstic, University of Nis, Serbia
Thomas D. Lagkas, The University of Sheffield International Faculty, CITY College - Thessaloniki, Greece
Hadi Larijani, Glasgow Caledonian University, UK
Lazaros Gkatzikis, KTH Royal Institute of Technology, Sweden
Hoang Le, Sandia National Laboratories, USA
Bertrand Le Gal, Institut Polytechnique de Bordeaux (IPB), France
Brian Lee, Software Research Institute, Ireland
Isaac Lera, Universitat de les Illes Balears, Spain
Keqin Li, State University of New York - New Paltz, USA
Wenzhong Li, Nanjing University, China
Jia-Chin Lin, National Central University, Taiwan, ROC

Dusan Radovic, TES Electronic Solutions GmbH - Stuttgart, Germany
Adib Rastegarnia, Purdue University, USA
Ustijana Rechkoska Shikoska, University for Information Science & Technology "St. Paul the Apostle" - Ohrid, Republic of Macedonia
José Renato Silva, Federal University of Rio de Janeiro (UFRJ), Brazil
Yenumula Reddy, Grambling State University, USA
Piotr Remlein, Poznan University of Technology, Poland
Eric Renault, Telecom SudParis, France
Lorayne Robertson, University of Ontario Institute of Technology, Canada
Pawel Rózycki, University of IT and Management (UITM), Poland
Danguole Rutkauskiene, Kaunas University of Technology, Lithuania
Zsolt Saffer, Budapest University of Technology and Economics (BME-HIT), Hungary
S. M. Sajid, National University of Emerging Sciences, Pakistan
Abheek Saha, Hughes Systique Corporation, USA
Ramiro Sámano Robles, Instituto de Telcomunicaçoes, Portugal
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
Benjamin Schiller, TU Darmstadt, Germany
Hans Schotten, University of Kaiserslautern, Germany
Motoyoshi Sekiya, Fujitsu Laboratories of America, USA
Sergei Semenov, Broadcom, Finland
Sandra Sendra Compte, University Polytechnic of Valencia, Spain
Dimitrios Serpanos, University of Patras, Greece
Michelle Sibilla, Paul Sabatier University Toulouse 3, France
Vilmos Simon, Budapest University of Technology and Economics, Hungary
Nicolas Sklavos, KNOSSOSnet Research Group - Technological Educational Institute of Western Greece, Greece
Celio Marcio Soares Ferreira, LinuxPlace, Brazil
Marco Spohn, Federal University of Fronteira Sul, Brazil
Keattisak Sripimanwat, National Science and Technology Development Agency (NSTDA), Thailand
Kostas Stamos, University of Patras, Greece
Mirjana Stojanovic, University of Novi Sad, Serbia
Lars Strand, Nofas Management, Norway
Maciej Szostak, Wroclaw University of Technology, Poland
Daniele Tafani, Dublin City University, Ireland
Yutaka Takahashi, Kyoto University, Japan
Yoshiaki Taniguchi, Kindai University, Japan
Vicente Traver Salcedo, Universitat Politècnica de València, Spain
Richard Trefler, University of Waterloo, Canada
Vassilis Triantafillou, Technological Educational Institute of Western Greece, Greece
Thrasyvoulos Tsiatsos, Aristotle University of Thessaloniki, Greece
Kazuya Tsukamoto, Kyushu Institute of Technology-Fukuoka, Japan
Masahiro Umehira, Ibaraki University, Japan
Guillaume Valadon, French Network and Information Security Agency, France
John Vardakas, Iquadrat Barcelona, Spain
Johanna Vartiainen, University of Oulu, Finland
Manos Varvarigos, University of Patras, Greece

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

Alex Mendes, Murilo Machado, and Helvio Freitas

Achieving a Desired Deterministic Upper Bounded PAPR Value Using a Fast Adaptive Clipping Algorithm

Lamarana Mamadou Diallo
CentraleSuplec/IETR/SCEE
Avenue de la Boulaie-CS 47601
35576 CESSON-SEVIGNE CEDEX
FRANCE
Mamadou-Lamarana.Diallo@supelec.fr

Jacques Palicot
CentraleSuplec/IETR/SCEE
Avenue de la Boulaie-CS 47601
35576 CESSON-SEVIGNE CEDEX
FRANCE
jacques.palicot@centralesupelec.fr

Faouzi Bader
CentraleSuplec/IETR/SCEE
Avenue de la Boulaie-CS 47601
35576 CESSON-SEVIGNE CEDEX
FRANCE
faouzi.bader@supelec.fr

***Abstract*—Orthogonal Frequency Division Multiplexing (OFDM) is the most commonly used multicarrier modulation in telecommunication systems due to the efficient use of the frequency resources and its robustness to multipath fading channels. However, as multicarrier signal in general, Peak-to-Average-Power Ratio (PAPR) is one of the major drawbacks of OFDM signals. Many works exist in the scientific literature on PAPR mitigation such as clipping methods, Tone Reservation based approaches, Partial Transmit Signals. However, in this paper we focus on clipping methods. This last is one of the most efficient adding signal techniques for PAPR reduction in terms of complexity. Nevertheless, clipping presents many drawbacks such as bit error rate degradation, out-of-band emission and mean power degradation. Adaptive clipping has been recently proposed in order to decrease these drawbacks. However, this approach is expensive in terms of numerical complexity, because an optimal threshold should be found for each OFDM symbol. This paper proposes a new approach to efficiently achieve the adaptive clipping, in terms of iterations number to find the optimal threshold. Theoretical analysis and simulation results validate the interest of this new clipping method.**

Keywords–OFDM, PAPR, CCDF, Clipping

I. INTRODUCTION

The Peak-to-Average Power Ratio (PAPR) is one of the main issues of the Orthogonal Frequency Division Multiplex (OFDM) signal. Many works [1][2][3] exist in the literature for PAPR mitigation. Clipping [4][5][6] method is an efficient technique for PAPR mitigation where the peak-canceling signal is computed by clipping the amplitudes of the signal that exceed a predefined threshold A. In practice, a normalized threshold $\rho = \frac{A^2}{P_{x_n}}$ is used, where $P_{\mathbf{x}_n}$ represents the mean power of the discrete signal $\mathbf{x}_n$ whose PAPR has to be reduced. It can be noted that, the normalized threshold defines the PAPR below which the signal is not clipped.

Due the large amplitude variations of the OFDM signals in the time domain, the instantaneous PAPR of each OFDM symbol highly depends on its content. Therefore, the instantaneous PAPR after Classical Clipping(CC) method [4] with a predefined normalized threshold also depends on its content. Then, the upper bounded PAPR of the clipped signal, at each value of its Complementary Cumulative Distribution Function (CCDF), increases when the CCDF decreases. This is illustrated by the left curve in Figure 1. Note that this is also the case for the original OFDM CCDF curve. That means there is no deterministic upper bounded PAPR for CC method. It is exactly what we target in this work. This deterministic value corresponds to the vertical solid blue line depicted in Figure 1. In practice, the suitable upper bounded PAPR of the signal

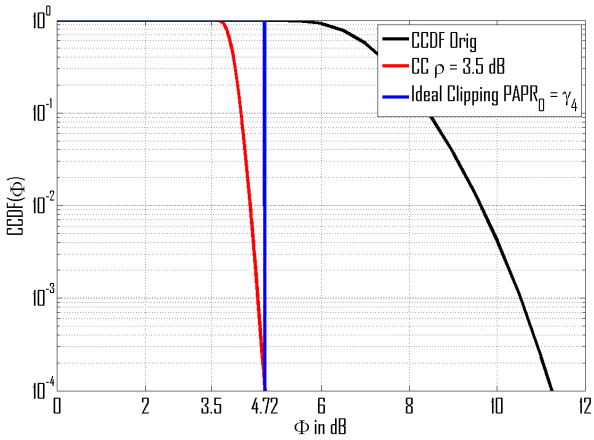


Figure 1. Scenario of CCDF curves of a classical clipping and Ideal Clipping.

for the Input Back Off (IBO) definition on the High Power Amplifier (HPA) is in general chosen at CCDF(Φ) close to zero (generally 10^{-4}). In this paper, this value is called the desired upper bounded PAPR and denoted as PAPR$_0$). Thus, in [7] the authors have shown that in CC techniques many OFDM symbols are either severely clipped or unnecessarily clipped with respect to this desired upper bounded PAPR. To illustrate this assertion, let us consider Figure 2 which is a zoom around 10^{-1} of the CCDF of the Figure 1. Note that our main objective is to have a PAPR clipping output about 4.72 dB (the vertical blue line). Therefore, all the symbols that have a PAPR value between 4.1 dB and 4.72 dB are clipped unnecessarily (see Δ_1 in figure 2). Besides this, all the symbols whose PAPR values are between 4.72 dB and 8.4 dB are severally clipped by the CC technique compared to ideal clipping (see indicated Δ_2 in figure 2). If we extend these considerations to all CCDF values, then we obtain the two areas of Figure 3:

- Area1: symbols are unnecessarily clipped

- Area2: Symbols are clipped more severely than necessary

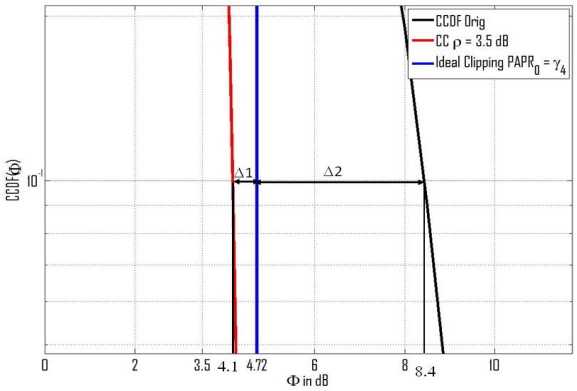


Figure 2. Zoom at CCDF=10^{-1} to illustrate symbols which are to much clipped by CC.

To avoid this drawbacks, the authors have proposed an Adaptive Clipping (AC) algorithm [7] in which the threshold is adapted to the content of each OFDM symbol and the desired upper bounded PAPR. Other adaptive clipping methods exist in the literature [8][9]. In [9], the authors proposed to adapt the normalized threshold ρ depending on the mapping constellation of the OFDM signal for a better compromise between PAPR reduction and BER degradation. In [8], the authors proposed an iterative clipping and filtering scheme [10] in which the computation of the amplitude threshold A from the predefined normalized threshold, is done at each iteration. This approach improves the performances on PAPR reduction but degrades more the signal. In contrast, in [7], the AC proposed approach and the classical clipping method [4] achieve same performance in terms of PAPR reduction. However, better bit error rate (BER), less out-of-band (OOB) emission and less mean power degradation are achieved. Nevertheless, the computational complexity of the proposed algorithm is high. In fact, from a predefined desired upper bounded PAPR (PAPR$_0$) and an initial normalized threshold ρ_0 = PAPR$_0$, an exhaustive search is performed to find the optimal threshold. For this purpose, having a predefined step $\epsilon > 0$, we check successively the values ρ_0, $\rho_0 - \epsilon$,...,$\rho_0 - k\epsilon$. In this context, the number of iterations to find the optimal threshold $\rho^{(\text{opt})}$ depends on the content of each OFDM symbol and ϵ. In this paper, we propose an efficient approach to compute $\rho^{(\text{opt})}$, which consists to adapt the step ϵ at each iteration. This technique is equivalent to clipping the signal iteratively by adapting A in function of PAPR$_0$ and the content of the clipped signal at the previous iteration. Therefore, we named this approach as Iterative Adaptive Clipping (IAC).

The paper is organized as follows. In Section II, the problem formulation and AC principle will be briefly presented. In Section III, we will present IAC approach and show that IAC method performs fewer iterations than AC approach to reach $\rho^{(\text{opt})}$. A comparative study in terms of signal degradation with the classical clipping will then be conducted in Section IV. The conclusions will be presented in Section V.

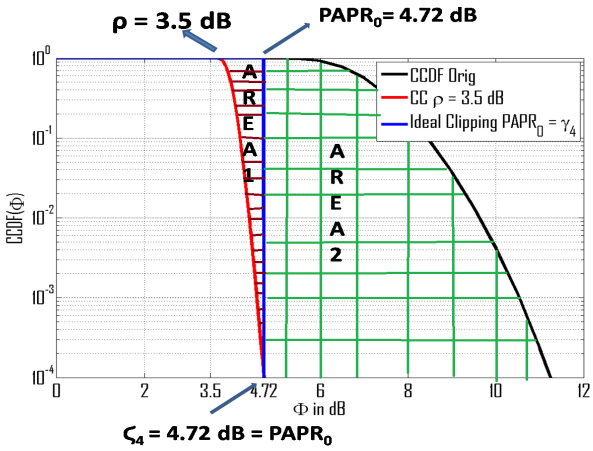


Figure 3. Scenario of CCDF curves of a classical clipping and Ideal Clipping.

II. ADAPTIVE CLIPPING ANALYSIS PRINCIPLE

Throughout this paper an OFDM signal $x_n(t)$ is given by the following equation

$$x(t) = \sum_{n=-\infty}^{+\infty} \sum_{m=0}^{M-1} X_{m,n} g(t - nT_u)\, e^{j2\pi mFt} \qquad (1)$$

where M means the total carriers, g is the window function of duration T_u, $F = \frac{1}{T_u}$ is the intercarrier space, mF the m^{th} frequency, and $X_{m,n}$ the symbol carried out by the m^{th} carrier at time nT_u.

In this paper, if $\mathbf{z}$ denotes a vector containing the time domain samples of the signal $z(t)$ in continuous time domain, its PAPR will be denoted by PAPR$_{\mathbf{z}}$. The positive scalar γ_e will represent in this paper the upper bounded PAPR of the signal at the CCDF value equal to 10^{-e} (e constant), i.e.,

$$\gamma_e = \max_{\Phi} \left\{ \text{CCDF}_{\mathbf{y}_n}(\Phi) \geq 10^{-e} \right\} \qquad (2)$$

where $\text{CCDF}_{\mathbf{y}_n}(\Phi) = \mathbb{P}rob[\text{PAPR}_{\mathbf{y}_n} \geq \Phi]$ and $\mathbf{y}_n$ is the signal after clipping. Let $\mathbf{x}_n = [x_{n,0}, \ldots, x_{n,NL-1}]^T$ be the vector containing the samples of the OFDM signal $x_n(t)$ oversampled by a factor L. The PAPR of $x_n(t)$ can be approximated from $\mathbf{x}_n$, as follows:

$$\text{PAPR}_{\mathbf{x}_n} = \frac{\max\limits_{m=0,\ldots,NL-1} \left\{ |x_{n,m}|^2 \right\}}{P_{\mathbf{x}_n}} \qquad (3)$$

where $P_{\mathbf{x}_n}$ is the mean power of the $\mathbf{x}_n$ signal before clipping.

The Classical Clipping (CC) proposed in [4] is one of the most popular clipping technique for PAPR reduction known in the literature. It is sometimes called hard clipping or soft clipping. To avoid any confusion, the (CC) name will be used in this paper. In [4], its effects on the performance of OFDM, including the power spectral density, the PAPR and BER are evaluated. The function-based clipping used for CC technique is defined as;

$$f(r, \text{A}) = \begin{cases} r, & r \leq \text{A} \\ A, & r > \text{A} \end{cases} . \qquad (4)$$

where A is the clipping threshold. From this equation, the PAPR of the output signal $\mathbf{y}_n$ after CC, if some samples of $\mathbf{x}_n$ are greater than the clipping threshold A, is given as follows:

$$\text{PAPR}_{\mathbf{y}_n} = \frac{A^2}{P_{\mathbf{y}_n}} \quad (5)$$

Given, $A = \left(10^{\frac{\rho}{20}}\right)\sqrt{P_{\mathbf{x}_n}}$, the PAPR of $\mathbf{y}_n$ can be rewritten as follows:

$$\begin{aligned} \text{PAPR}_{\mathbf{y}_n} &= \left(10^{\frac{\rho}{10}}\right)\left(\frac{P_{\mathbf{x}_n}}{P_{\mathbf{y}_n}}\right) \\ \text{then PAPR}_{\mathbf{y}_n}(\text{ in dB }) &\geq \rho(\text{ in dB}) \end{aligned} \quad (6)$$

Therefore, it can be noticed that $\gamma_e \geq \rho$ for any $e \geq 0$. So γ_e increases when e increases. In practice, the desired γ_e for IBO parameterization of the HPA is generally chosen at CCDF value equal to 10^{-4}, i.e., γ_4. Then, we may remark that by using CC method many OFDM symbols are clipped more severely than necessary or unnecessarily clipped with respect to γ_4 [7]. Figure 3 shows the domains representing the set of OFDM symbols which are clipped more severely than necessary (AREA2) or unnecessarily clipped (AREA1) for a CC with $\rho = 3.5$ dB, in respect to Ideal Clipping (see vertical blue line of Figure 3), for the same upper bounded PAPR at CCDF value equal to 10^{-4} (γ_4). The vertical blue line represents the ideal clipping CCDF for $\text{PAPR}_0 = \gamma_4$, which corresponds to the deterministic desired upper bounded PAPR. It is obvious that the output upper bounded PAPR of such ideal clipping is constant at any value of the CCDF. In [7], this ideal clipping has been approached by AC. The theoretical analysis and simulation results achieved by the authors have demonstrated that AC method outperforms CC in terms of signal degradation having the same performance on PAPR reduction.

The AC method consist to adapt the normalized threshold ρ_n for each OFDM symbol $\mathbf{x}_n$ which we want to clip with respect to the desired upper bounded PAPR value PAPR_0 by solving the following equation

$$\text{PAPR}_0 = \gamma_4 = \left(10^{\frac{\rho_n}{10}}\right)\left(\frac{P_{\mathbf{x}_n}}{P_{\mathbf{y}_n}}\right), \quad (7)$$

where $P_{\mathbf{y}_n}$ is the mean power of the clipped signal $\mathbf{y}_n$ with ρ_n.

From (7), it can be noticed that $P_{\mathbf{y}_n}$ depends on the unknown parameter ρ_n. In [7], an exhaustive research in $[0, \gamma_4]$ is proposed to solve (7). Having $\epsilon > 0$, the authors proposed to check successively $\rho_0 = \gamma_4$, $\rho_1 = \rho_0 - \epsilon$, ... , $\rho_m = \rho_{m-1} - \epsilon$,..., to reach $\rho^{(\text{opt})}$ which satisfies

$$(\text{PAPR}_{\mathbf{y}_n} - \gamma_4) \leq \delta, \quad (8)$$

where $\delta > 0$ is a satisfactory residual error. In other words, if the algorithm has performed m iterations, then $\rho^{(\text{opt})} = \gamma_4 - m\epsilon$. Note that, in AC, the step (ϵ) is constant at each iteration. Therefore, the number of necessary iterations to find $\rho^{(\text{opt})}$ depends on the content of each OFDM symbol,γ_4, and ϵ. A less complex approach is proposed in this paper. The main idea is to find $\rho^{(\text{opt})}$ in the interval $[0, \gamma_4]$ with few iterations. For this purpose, we propose, for each OFDM symbol, to adapt the step ϵ at each iteration in order to increase the convergence rate towards $\rho^{(\text{opt})}$. The following section presents the description of the IAC proposed approach.

III. ITERATIVE ADAPTIVE CLIPPING APPROACH

In this section, we present the IAC proposed method and theoretical analysis of its performances in terms of PAPR reduction. Theoretical comparison with AC in terms of convergence speed will be also presented.

We denote by $f(., A)$ the CC function, see (4), used in [4]. Having $\delta > 0$, IAC approach consists of searching the normalized threshold $\rho^{(\text{opt})}$ which satisfies (8). To find this threshold we check successively $\rho_0 = \text{PAPR}_0$, $\rho_1 = \rho_0 - \epsilon_1$,...,$\rho_m = \rho_{m-1} - \epsilon_m$. ϵ_m is the step between ρ_{m-1} and ρ_m. In others words it is the step which will give the ρ_m threshold, which will be used to clipp $\mathbf{y}_n^{(m-1)}$. $\mathbf{y}_n^{(m-1)}$ being the ouput signal after the $(m-1)^{\text{th}}$ clipping iteration. Note that ϵ_m is not constant and should depend on the content of each OFDM symbol and its clipped version at the previous iterations. Therefore, we defined the step ϵ_m at the m^{th} iteration as

$$\epsilon_m = 10\text{Log}_{10}\left(\frac{P_{\mathbf{y}_n^{(m-2)}}}{P_{\mathbf{y}_n^{(m-1)}}}\right), \quad (9)$$

with the notation $P_{\mathbf{y}_n^{(-1)}} = P_{\mathbf{x}_n}$ at the first iteration. The flow chart of the IAC approach is given in Figure 4.

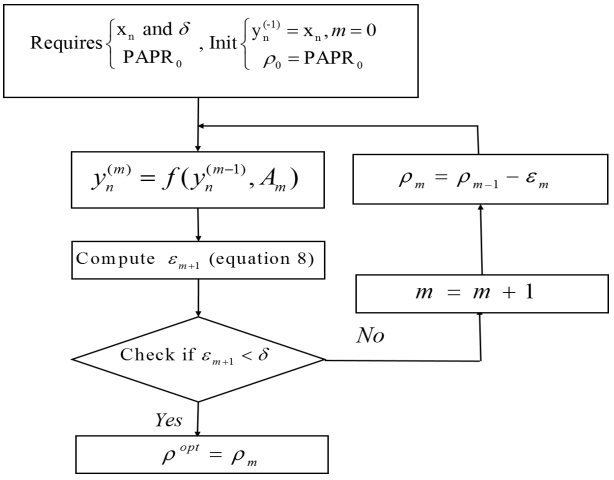


Figure 4. Flow chart of the IAC approach.

The amplitude threshold A_m at the m^{th} iteration can be expressed from the corresponding normalized threshold ρ_m as follows:

$$\begin{aligned} A_m &= 10^{\frac{\rho_m}{20}}\sqrt{P_{\mathbf{x}_n}} \\ &= \left(10^{\frac{\rho_0 - \epsilon_1 - \ldots - \epsilon_m}{20}}\right)\sqrt{P_{\mathbf{x}_n}} \\ &= \left(10^{\frac{\text{PAPR}_0}{20}}\right)\left(10^{\frac{-\sum_{l=1}^{m}\epsilon_l}{20}}\right)\sqrt{P_{\mathbf{x}_n}} \\ &= \left(10^{\frac{\text{PAPR}_0}{20}}\right)\left(\prod_{l=1}^{m} 10^{\frac{-\epsilon_l}{20}}\right)\sqrt{P_{\mathbf{x}_n}} \end{aligned} \quad (10)$$

Then, from (9) we obtain the following expression after some

derivation

$$A_m = \left(10^{\frac{\text{PAPR}_0}{20}}\right)\left(\prod_{l=1}^{m}\sqrt{\frac{P_{\mathbf{y}_n^{(l-1)}}}{P_{\mathbf{y}_n^{(l-2)}}}}\right)\sqrt{P_{\mathbf{y}_n}} = \left(10^{\frac{\text{PAPR}_0}{20}}\right)\sqrt{P_{\mathbf{y}_n^{(m-1)}}} \quad (11)$$

Then, by substituting (11) in (5) the PAPR of the clipped signal at the m^{th} iteration satisfies the following expression:

$$\text{PAPR}_{[\mathbf{y}_n^{(m)}]} - \text{PAPR}_0 = \epsilon_{m+1}. \quad (12)$$

Therefore, if we define $\epsilon_{m+1} \leq \delta$ as the criteria for stopping IAC at the m^{th} iteration, then, for each OFDM symbol the PAPR of the output signal after PAPR reduction by IAC is less than $\text{PAPR}_0 + \delta$. So, the CCDF curve of the IAC will approach the ideal clipping and give the desired deterministic upper bounded PAPR. The following Algorithm 1 describes the IAC proposed technique.

Algorithm 1 IAC algorithm

Require: $\mathbf{x}_n$ input OFDM signal , $\delta > 0$ and PAPR_0
Ensure: $\mathbf{y_n}$ output signal
  $m \leftarrow 0$
  $\epsilon_m \leftarrow 1$
  $\mathbf{y}_n^{(-1)} \leftarrow \mathbf{x}_n$
  while $\left(\text{PAPR}_{\mathbf{y}_n^{(m)}} - \text{PAPR}_0\right) = \epsilon_m \geq \delta$ **do**
    $m \leftarrow m + 1$
    Compute A_m from equation 11
    $\mathbf{y}_n^{(m)} \leftarrow f(\mathbf{y}_n^{(m-1)}, A_m)$
  end while

For convergence speed comparison with AC, we can remark that, at each iteration, AC and IAC algorithms have almost the same numerical complexity. Therefore, convergence speed comparison will be achieved by comparing the number of iterations performed by these algorithms for each OFDM symbol.

For each OFDM symbol $\mathbf{x}_n$ and $\delta > 0$, let $N_{\mathbf{x}_n,1}, N_{\mathbf{x}_n,2}$ be the number of iterations performed by AC and IAC to converge towards $\rho^{(\text{opt})}$, respectively. So, $\rho^{(\text{opt})}$ is approximated by $\rho_{N_{\mathbf{x}_n,1}} = \text{PAPR}_0 - N_{\mathbf{x}_n,1}\epsilon$ and $\rho_{N_{x_n,2}} = \text{PAPR}_0 - \sum_{l=1}^{N_{\mathbf{x}_n,2}} \epsilon_l$ in AC and IAC, respectively. Let's define the average step for the IAC as,

$$\epsilon_{\mathbf{x}_n} = \frac{1}{N_{\mathbf{x}_n,2}} \sum_{l=1}^{N_{\mathbf{x}_n,2}} \epsilon_l. \quad (13)$$

Then, for each OFDM symbol $\mathbf{x}_n$, the number of iterations performed by IAC to find $\rho^{(\text{opt})}$ is equal to the number of iterations performed by AC when the step is equal to $\epsilon_{\mathbf{x}_n}$. In fact, from (6) the PAPR of the output signal at the m^{th} in AC with the step $\epsilon_{\mathbf{x}_n}$ can be expressed as follows

$$\text{PAPR}_{[\mathbf{y}_n^{(m)}]} = \text{PAPR}_0 - m\epsilon_{\mathbf{x}_n} + 10\text{Log}_{10}\left(\frac{P_{\mathbf{x}_n}}{P_{\mathbf{y}_n^m}}\right). \quad (14)$$

After few derivations and by using (13), we obtain

$$\text{PAPR}_{[\mathbf{y}_n^{(m)}]} - \text{PAPR}_0 = 10\text{Log}_{10}\left[\left(\frac{P_{\mathbf{y}_n^{(m-1)}}}{P_{\mathbf{x}_n}}\right)^{\frac{m}{N_{\mathbf{x}_n,2}}}\frac{P_{\mathbf{x}_n}}{P_{\mathbf{y}_n^{(m)}}}\right]$$

Therefore, since the number of iterations performed by IAC to compute the normalized threshold for the OFDM symbol $\mathbf{x}_n$ is $N_{\mathbf{x}_n,2}$ we remark that

$$\begin{cases} \left(\text{PAPR}_{[\mathbf{y}_n^{(m)}]} - \text{PAPR}_0\right) \geq 10\text{Log}_{10}\left[\frac{P_{\mathbf{y}_n^{(m-1)}}}{P_{\mathbf{y}_n^{(m)}}}\right] \\ \qquad\qquad \geq \epsilon_m > \epsilon \text{ If } m < N_{\mathbf{x}_n,2} \\ \left(\text{PAPR}_{[\mathbf{y}_n^{(m)}]} - \text{PAPR}_0\right) = \text{Log}_{10}\left[\frac{P_{\mathbf{y}_n^{(N_{\mathbf{x}_n,2}-1)}}}{P_{\mathbf{y}_n^{(N_{\mathbf{x}_n,2})}}}\right] \\ \qquad\qquad = \epsilon_{N_{\mathbf{x}_n,2}+1} < \epsilon \text{ If } m = N_{\mathbf{x}_n,2} \end{cases}$$

which proves that, for each $\mathbf{x}_n$ the number of iterations performed by IAC is equal to the number of iterations performed by AC in which the step is equal to $\epsilon_{\mathbf{x}_n}$. Thus, for each OFDM symbol, the comparison between $N_{\mathbf{x}_n,1}$ and $N_{\mathbf{x}_n,2}$ can be achieved by comparing $\epsilon_{\mathbf{x}_n}$ and ϵ. However, since $\mathbf{x}_n$ is a random signal we will compare IAC and AC by comparing the mean of number of iterations required for each algorithms. This is equivalent to compare $\mathbb{E}[\epsilon_{\mathbf{x}_n}]$ defined in (15) and ϵ (the constant step in AC) as,

$$\mathbb{E}[\epsilon_{\mathbf{x}_n}] \simeq \frac{1}{P_{\mathbf{x}_n}} \sum_{m=0}^{N_2} \int_0^{+\infty} f(r, A_m)p(r)\text{d}r \quad (15)$$

where $p(r)$ is the probability density function of the amplitudes of the signal OFDM signal and $N_2(resp N_1)$ represent the mean of the number of iterations performed by IAC (resp AC) over a great number of K OFDM symbols.

$$N_i = \frac{1}{K}\sum_{n=0}^{K} N_{\mathbf{x}_n,i}, i = 1, 2 \quad (16)$$

After some computations [6] we obtain,

$$\mathbb{E}[\epsilon_{\mathbf{x}_n}] = \frac{1}{P_{\mathbf{x}_n}} \sum_{m=0}^{N_2} \left(1 - e^{\frac{-A_m^2}{P_{\mathbf{x}_n}}}\right) \quad (17)$$

In [7], in order to obtain the desired upper bounded PAPR equal to $\text{PAPR}_0 + \delta$, the step ϵ must be chosen less or equal to δ. It is clear that the number of iterations increases when ϵ decreases. Thus, the optimal step in AC is $\epsilon = \delta$. From (17) and the fact that in IAC $\epsilon_m > \epsilon$ if $m < N_{\mathbf{x}_n,2}$, IAC converges more quickly than AC if $\epsilon_1 \geq \epsilon$. Therefore, from (9) we can deduce that, for each PAPR_0 $N_1 \geq N_2$ if and only if

$$\epsilon_1 = 10\text{Log}_{10}\left(\frac{1}{1 - e^{-\text{PAPR}_0}}\right) \geq \epsilon.$$

After some derivations, we can conclude that:

$$N_1 \geq N_2 \text{ If and only if } \text{PAPR}_0 \leq \ln\left(\frac{10^{\frac{\epsilon}{10}}}{10^{\frac{\epsilon}{10}} - 1}\right) \quad (18)$$

IV. SIMULATION RESULTS

The performance of the proposed IAC and the CC method are analyzed under same PAPR reduction, i.e., $\text{PAPR}_0 = \gamma_4 - \delta$, and compared in terms of signal degradation and convergence speed. The simulations are performed for an OFDM signal

with 16-QAM modulation in which $M = 64$, an oversampling factor $L = 4$.

Figure 5 confirms that the CCDF curves of the IAC approximate the ideal clipping CCDF curve. Beisdes, the IAC and CC method achieve the same upper bounded PAPR value at CCDF equal to 10^{-4}. It can be also noticed from depicted results, that IAC method reach a deterministic upper bounded PAPR.

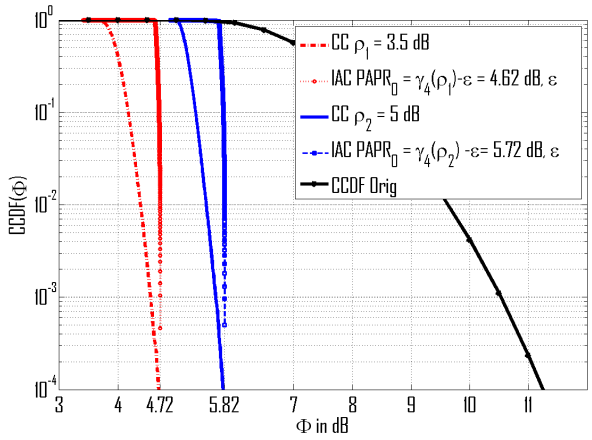


Figure 5. Performance of IAC in terms of PAPR reduction for different thresholds $\rho_1 = 3.5$dB and $\rho_2 = 5$dB

In the following, the IAC is compared with CC in terms of BER degradation.

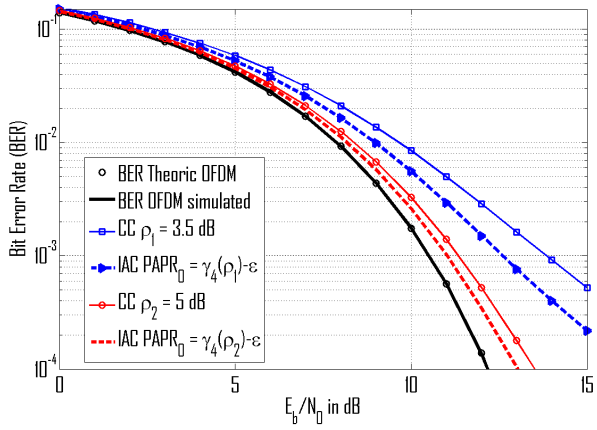


Figure 6. Comparison of CC and IAC in terms of BER degradation for $\rho = 3.5$dB.

Results depicted in Figure 6 show that IAC outperforms CC in terms of BER degradation. The obtained gain at 10^{-4} of BER, is greater than 1 dB. This result confirms the theoretical analysis undertaken in [7] where the authors have shown that in CC many OFDM symbols are clipped more severely (see AREA 2 in Figure 3) than necessary or unnecessarily (see AREA 1 in Figure 3) with respect to γ_4.

The performances in terms of Mean Power degradation and adjacent channels pollution which is due to the effect of the OOB components, are depicted in Figure 7.

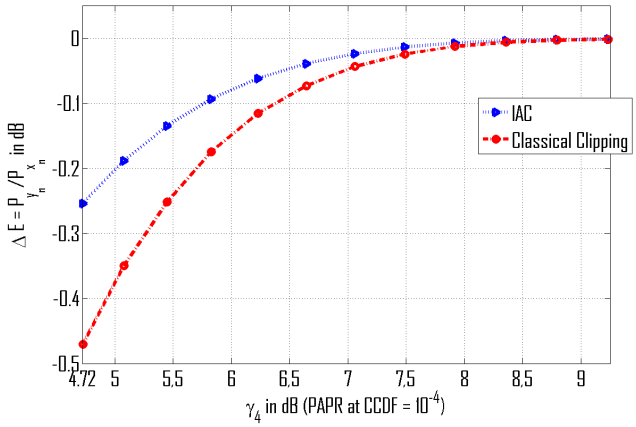


Figure 7. Comparison of CC and IAC in terms of Mean Power degradation for $\rho = 3.5$dB and a PAPR at CCDF 10^{-4}

.

From the simulation results depicted in Figure (7), it can be noticed that IAC degrades less the Mean Power of the clipped signal than the CC for the same PAPR performance reduction at a CCDF $\leq 10^{-4}$. For example, for $\gamma_4 = 4.72$dB, $\Delta E = -0.47$dB in CC method and $\Delta E = -0.25$dB in proposed IAC approach.

Figure (8) represents the Power Spectrum Density (PSD) of both OFDM signal before PAPR reduction and after PAPR reduction by IAC and CC respectively.

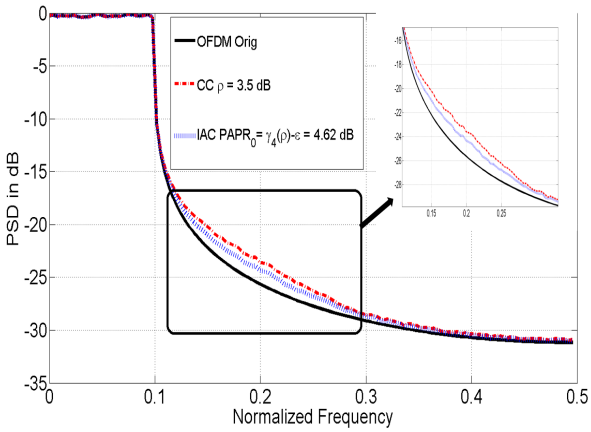


Figure 8. Comparison of the OFDM PSD using IAC and CC for threshold $\rho = 3.5$dB

Similar as in Figure 7, for BER degradation and mean power variation, Figure 8 shows that IAC pollutes less the adjacent channels than CC when $\text{PAPR}_0 = \gamma_4 - \epsilon$.

As a general conclusion, obtained results in terms of signal degradation confirm that when $\text{PAPR}_0 = \gamma_4 - \epsilon$, IAC

 ISBN: 978-1-61208-411-4

degrades less the signal than CC method (see Figure 6, 7, 8).

In the following, we compare N_1 and N_2 defined by

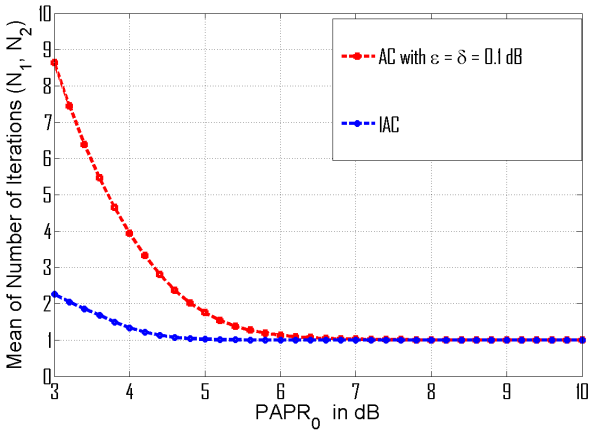


Figure 9. Mean of number of iterations performed by IAC and AC for each OFDM symbol in function of $PAPR_0$

equation (16) by simulation with $K = 10^4$. Figure 9 shows that IAC method converges more quickly than AC method, for instance, when $\gamma_4 = 3$dB, and $N_1 \simeq 4N_2$.
Obtained results confirms our theoretical analysis undertaken in Section III (see equation (18)). In fact, from Figure (9), it can be remarked that $N_1 \geq N_2$ when $\gamma_4 \leq 6$ which is coherent with equation (18) (with $\epsilon = 0.1$dB $\Rightarrow 10\text{Log}_{10}\left[\text{Log}\left(\frac{10^{\frac{0.1}{10}}}{10^{\frac{0.1}{10}}-1}\right)\right] = 5.77\text{dB} \simeq 6\text{dB}$).

V. CONCLUSION

In this paper, a new method for approximating the normalized adapted threshold for the adaptive clipping is presented. The theoretical analysis and simulation results achieved in this paper show that this approach converges more quickly than the one based on exhaustive research with a constant step. This approach outperforms also CC in terms of signal degradation, with the same performances in terms PAPR reduction. Furthermore, IAC gives a deterministic desired upper bounded PAPR which is very important for IBO definition on high power amplifiers (HPA). Our future work will focus on the extension of proposed work to other clipping functions as deep clipping and smooth clipping combined with Out Of Band noise suppression approcahes.

ACKNOWLEDGMENT

Part of this work is supported by the project ACCENT5 (Advanced Waveforms, MAC Design and Dynamic Radio Resource Allocation for D2D in 5G Wireless Networks) funded by the French national research agency with grant agreement code: ANR-14-CE28-0026-02.

REFERENCES

[1] Y. Louet and J. Palicot, “A classification of methods for efficient power amplification of signals,” in Annals of telecommunications, vol. 63, 2008, pp. 351–368.

[2] J. Tellado-Mourelo, “Peak to average power reduction for multicarrier modulation,” Ph.D. dissertation, Standford University, 1999.

[3] S. Zabre, J. Palicot, Y. Louet, and C. Lereau, “SOCP approach for OFDM peak-to-average power ratio reduction in the signal adding context,” in Proc of Signal Processing and Information Technology, 2006 IEEE International Symposium on, 2006, pp. 834–839.

[4] X. Li and L. Cimini, “Effects of clipping and filtering on the performance of OFDM,” in Vehicular Technology Conference, 1997, IEEE 47th, vol. 3, 1997, pp. 1634–1638 vol.3.

[5] S. Kimura, T. Nakamura, M. Saito, and M. Okada, “Par reduction for OFDM signals based on deep clipping,” in Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on, 2008, pp. 911–916.

[6] D. Guel, “Etude de nouvelles techniques de réduction de ”facteur de crête” compatibilité descendante pour les systemes multiporteuses.” Ph.D. dissertation, Université de Rennes 1 (French), 25 Novembre 2009.

[7] L. M. Diallo and J. Palicot, “Adaptive clipping for a deterministic PAPR,” (invited paper) ICTRS,in Proceedings Of the Third International Conference on Telecommunications and Remote Sensing, June 2014, Luxembourg.

[8] Y. K. Byuong Moo Lee, “An adaptive clipping and filtering technique for papr reduction of OFDM signals,” Circuit, Systems and Signal Processing, vol. 32, 2013, pp. 1335–1349.

[9] H. J. Kim, S. C. Cho, H. S. Oh, and J. M. Ahn, “Adaptive clipping technique for reducing PAPR on OFDM systems,” in Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th, vol. 3, Oct 2003, pp. 1478–1481 Vol.3.

[10] J. Armstrong, “Peak-to-average power reduction for OFDM by repeated clipping and frequency domain filtering,” Electronics Letters, vol. 38, no. 5, Feb 2002, pp. 246–247.

 ISBN: 978-1-61208-411-4

A Robust Bit Modification Audio Steganography for Covert Communication

Kaliappan Gopalan
Department of Electrical and Computer Engineering
Purdue University Calumet
Hammond, U.S.A.
e-mail: gopalan@purduecal.edu

Jiajun Fu
Department of Electrical and Computer Engineering
Purdue University Calumet
Hammond, U.S.A.
e-mail: fu77@purduecal.edu

***Abstract*—Sample bit modification for data embedding on a cover audio signal has been shown as a viable technique for steganography and watermarking. Depending on the sample bit index chosen for carrying the embedded data, there is a tradeoff between viability of the data in the presence of noise, robustness and imperceptibility. Although a high threshold of audio samples can carry data at higher bit indices thereby raising robustness, it can be susceptible to noise even at low levels and, with sample amplitudes changed significantly, embedding becomes conspicuous, both detrimental for covert or secure communication. In this paper, modification of the high threshold sample embedding is shown to increase noise immunity with correct data retrieval at a lower payload, but without sacrificing indiscernibility. Experimental results using a noise-free utterance (from a corpus of read speech) and a noisy utterance (between air traffic controllers and pilots) show zero to low bit error rate of hidden data recovery at added noise levels of 50 decibels of signal-to-noise ratio.**

Keywords— Audio steganography; data embedding; bit modification; perceptual quality measure; noise robustness; stego audio.

I. INTRODUCTION

Secure and covert communication using unsecured network relies on steganographic techniques employing audio, image and video as host or cover carriers. Applications of such secure communication abound in battlefield data transmission and civilian transmission of banking, medical and employment data, to name a few. Steganography in general, and audio steganography more specifically, can supplement and enhance encrypted digital data for added security and privacy.

While the challenge of meeting all the key criteria of high payload, low or no perceptibility of embedding and high data integrity in the presence of noise is hard, applications with different requirements can readily be satisfied with tradeoff in one or more criteria. Watermarking of speech for copyright protection or authenticity verification, for example, may not need as much payload as for transmitting confidential medical data. Additionally, music copyright and/or transmission requires high level of indiscernibility. Covert communication may need to carry a reasonably high volume of information with little noticeability of the presence of embedding. Efficacy of techniques of data hiding can, therefore, be different with varying degrees of fulfilling the criteria. Additionally, use of the original host, or cover, audio signal for retrieving the embedded information may not be a limitation in watermarking applications; for covert communication, however, this type of escrow detection of hidden data may be an impediment requiring the use of the same host signal at the receiver and transmitter. It also may cause suspicion about the audio signal hiding information. Oblivious retrieval, on the other hand, needs some property of the host signal to remain the same in the stego, or data-embedded signal. A generally used invariant property is the psychoacoustic masking phenomenon of the human auditory system that renders spectral changes in an audio signal that are below its global masking threshold imperceptible. If the embedding procedure leaves the resulting spectral changes below the masking threshold, the stego becomes indiscernible from the host. In addition, the same masking threshold can be used at the receiver to retrieve the embedded data. Based on these key advantages, a number of techniques have been developed for audio steganography with oblivious detection [1]-[4].

The paper is organized as follows. Section II provides a brief review of audio sample bit modification for embedding. In Section III the proposed bit modification technique is described. Experimental results observed and a discussion of these results are given in Section IV. Conclusions drawn from the work form Section V.

II. AUDIO STEGANOGRAPHY EMPLOYING TIME DOMAIN SAMPLE BIT MODIFICATION

An alternative to hiding data in the spectral domain of host audio that exploits the auditory masking property of human perception is to alter time-domain samples in according with the data. Time-domain sample modification maintains imperceptibility if small changes are made to a few samples that are in the neighborhood of relatively large samples, for example. An early sample modification technique replaced the least significant bit (lsb) of each of a selected set of host audio samples with the data to be embedded. Such a simple technique, clearly, is susceptible to loss of data due to noise and also to illegal removal or replacement of the lsb. Several higher order bit modification techniques carry data on samples that are large enough but at bit indices that contribute to relatively small changes so that audibility of embedding is reduced. While lower bit indices generally cause less noticeability of embedding with higher

 ISBN: 978-1-61208-411-4

payload, it is also more susceptible to noise [5]-[7]. In this paper, we report an imperceptible bit modification steganography that can recover hidden data in the presence of noise on the stego.

III. SAMPLE BIT MODIFICATION AT SIGNIFICANT SAMPLES

Employing high bit indices for carrying hidden data can alleviate noticeability of modification if the samples are large in amplitude and the modified bit is relatively small. While this may reduce payload for a given host audio – due to non-availability of a large number of high amplitude samples – it can help mask auditory perception and contribute to higher noise robustness. With this premise, the following bit modification procedure was carried out on a noise-free and a noisy audio signal, as an extension to previously reported bit modification steganography [8].

Samples of a given host audio signal are selected for carrying hidden data based on a threshold M, where $M = 2^{l1} + 2^{l2} + 2^{l3}$, so that only amplitudes a that satisfy $|a| \geq M$ are used for modification. For a 16-bit audio with full dynamic range, a typical threshold can use l1 = 10, l2 = 11 and l3 = 12 (with LSB at index 1) so that $M = 3584$; hence, only samples with magnitudes of at least 3584 are considered potential samples available for bit modification.

To reduce the significance of change due to one of the 16 bits modified in the set {S}, sample bit $k < l1$, the smallest index used for the threshold, is used for modification in accordance with data to be hidden if $\frac{2^k}{M} \leq r$.

This criterion ensures that the modified sample is different from the original host audio by no more than $100r$ %. By a choice of r, this empirical rule can result in minimal changes in stego while affording different higher order bit indices for embedding in larger sample values. Although a large bit index k may raise data robustness to noise, it can also cause noticeable change in spectrogram and audibility, both resulting in conspicuousness of embedding. A reasonable choice for the index k is, therefore, below the lowest threshold bit index $l1$, in general. Test results are shown in the following section for different values of k.

IV. EXPERIMENTAL RESULTS

The first test used a noise-free utterance (from the corpus of phonemically and lexically transcribed speech of American English speakers) available at a sampling rate of 16000 Hz. Using a threshold of $M = 3584$, i.e., with l1 = 10, l2 = 11 and l3 = 12, effect of modifying different bit indices of samples satisfying the threshold was studied for different levels of noise added to the stego. The host audio was windowed into 320 samples (20 ms) of non-overlapping segments. Only those segments that had a significant number of potential amplitudes (at least 10) were considered for carrying hidden data. To increase data robustness in the presence of noise, if a frame had at least 10 potential samples, each of these samples was modified at its k^{th} bit with the same single bit of data (or, data bit exclusive-ORed with a key) to be embedded. By using a majority of the 10 (or more) recovered bits, probability of correct bit recovery was increased at the cost of reduced payload. As an example, Figure 1 shows the spectrograms of the original (host) audio and the stego carrying 71 bits of data in each of the 71 frames. The host with 51544 samples had 161 frames with 71 frames having 10 or more samples that were larger than the threshold of $M = 3584$. Each of the first 10 significant samples in a frame was modified at its 7th bit (lsb = 1) with the same data bit. The data bit index value corresponded to the embedded frame index – frame 16 that satisfied both the threshold and the number of samples, for example, carried data bit 16 in all of its 10 or more samples. Thus, with 71 frames of the host audio, the stego carried 710 bits with 71 bits of data.

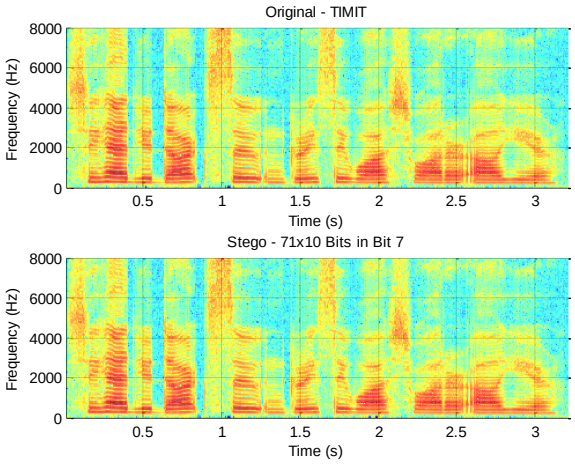


Figure 1. Spectrograms of original (noise-free) host (top) and stego carrying 71x10 bits at sample index 7 (lsb = 1)

Retrieval of the data bits was carried out by first segmenting the stego audio into samples of 320 and determining those samples that were above the threshold of 3584 in magnitude. If a segment had at least 10 such samples, then the k^{th} bit of each of the first 10 of these samples was obtained (with the same key as the one used for embedding). A majority of the 10 recovered bits was considered the correct embedded bit. With this procedure, all of the 71 embedded bits were correctly recovered.

To study the robustness of data with noise, zero-mean Gaussian noise at various levels of signal-to-noise ratio (SNR) was added to the stego. Data-retrieval from the noise-added stego was proceeded in the same manner as above. If the level of noise was such that the threshold was unaffected, the samples in which a bit in each was embedded remained the same; the noise, however, could have affected the k^{th} bit in some cases. By a majority voting of the recovered bits from the first 10 samples of each embedded frame, error due to noise was reduced. Figure 2 shows the original host audio and the noise-added stego at 50 dB of SNR. With majority voting, all 71 bits embedded in the stego were correctly recovered.

 ISBN: 978-1-61208-411-4

As the noise level was increased, either more samples were affected at the k^{th} bit of embedded samples, or worse, the noise altered the embedded samples so that threshold was not satisfied at the same frames as those used for embedding; both cases led to errors in data retrieval. Changing the bit index k for sample modification of the host, similarly, caused errors with lower levels of noise as k was decreased. Table I shows the data bit error rate (BER) as a function of modified bit index k and SNR. Each row corresponds to the maximum SNR for the k^{th} bit used for embedding. As the bit index k was reduced, noise tolerance became smaller and BER increased, although a BER of zero was achieved for the stego without any noise.

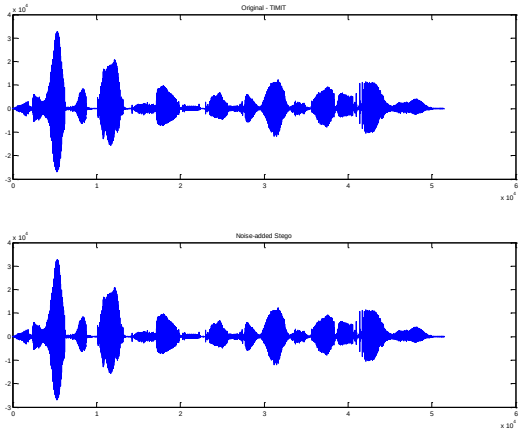


Figure 2. Waveforms of original (host) audio (top) and the noise-added stego carrying a total of 710 bits.

TABLE I. EMBEDDED BIT INDEX VS. NOISE VS. BER

k	SNR	BER, %
9	50	0
8	50	0
7	50	0
6	50	1.4085

All cases correspond to the same threshold of 3584 and 71 bits of data with each bit repeated 10 times in a frame with 10 or more samples above the threshold.

In the second test, a noisy audio was used as a practical example of host to carry hidden information. This audio from the Greenflag database of communication between fighter aircraft pilots and their air traffic controllers has 80150 samples obtained at the rate of 8000 per second. With 160 samples (20 ms) per frame, there were 500 frames and 302 of these frames satisfied the same threshold of 3584 with 20 or more samples. Choosing to repeat the same bit 20 times (the first 20 in each embeddable frame), all 302 bits were recovered from the total of 6040 bits. Correct data recovery was also achieved with noise at 50 dB SNR added to the stego. Figure 3 shows the spectrograms of the original host audio and the noise-added stego audio carrying data at sample bit index 7. At higher levels of noise, BER started to show up. Similar results were observed for embedding indices of 8 and 9, again with noise added at 50 dB or higher SNR.

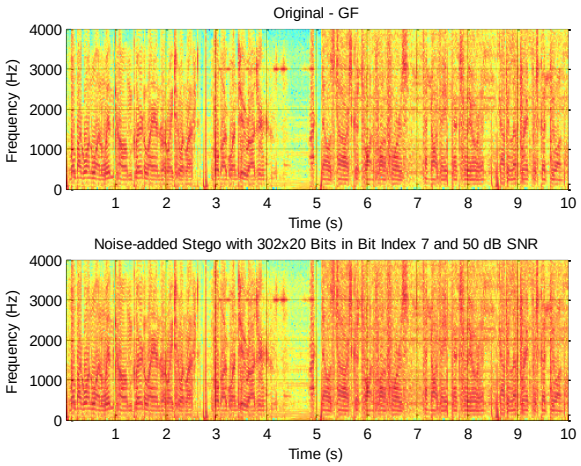


Figure 3. Spectrograms of host (noisy) audio (top) and noise-added stego carrying 302x20 bits at sample index 7

When the repetition rate of embedding the same bit in a frame was reduced to 10, there were 335 frames available for a data payload of 335 bits. At this slightly increased payload, one to three bits were incorrectly recovered at the noise level of 50 dB for a BER of 0.2985 to 0.8955, while no error resulted in the absence of added noise. This shows that the majority voting contributes to correct data recovery when noise is present in the stego. Similar results were observed at other indices for k, with BER increasing with noise level at lower indices.

From the two examples of audio considered, we may observe that a noise-free host audio is likely to have fewer samples satisfying a large threshold; payload, consequently, is reduced. A more realistic host audio with ambient and other type of noise, on the other hand, may have a high number of samples that can be modified with data without causing any perceptual or other difference.

V. CONCLUSION

An improved audio steganography employing time-domain sample bit modification at high bit indices has been proposed. Results observed on a clean and a noisy host audio signal show the viability of the technique in imperceptible embedding, and oblivious and error-free retrieval of the embedded data. By using a higher bit index of selected samples, and with a majority voting, hidden data bits can be correctly extracted even in the presence of added noise. The tradeoff for robust data recovery is low payload. The proposed method may be suitable for covert

 ISBN: 978-1-61208-411-4

communication of battlefield information or for secure transmission of medical and other data. Based on the imperceptibility of embedding, audio watermarking and authentication can also use this method.

REFERENCES

[1] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data embedding and watermarking technologies," Proc. IEEE, Vol. 86, June 1998, pp. 1064-1087.

[2] J. F. Tilki and A. A. Beex, "Encoding a Hidden Auxiliary Channel onto a Digital Audio Signal Using Psychoacoustic Masking , IEEE Southeastcon 97, April 1997, pp. 331-333.

[3] K. Gopalan, "Audio Steganography Using Bit Modification," Proc. of the IEEE 2003 International Conference on Multimedia and Exposition (ICME 2003), July 2003, pp. I-629-632.

[4] N. Cvejic and T. Seppanen, "Increasing robustness of LSB audio steganography by Reduced Distortion LSB Coding," Journal of Universal Computer Science, vol. 11, no. 1, pp 56-65, 2005.

[5] K. Gopalan and Qidong Shi, "Audio Steganography using Bit Modification – A Tradeoff on Perceptibility and Data Robustness for Large Payload Audio Embedding," Proc. of the 19th International Conference on Computer Communications and Networks (ICCCN 2010) -- Workshop on Multimedia Computing and Communications, Zurich, Switzerland, Aug. 2010, pp. 1-6.

[6] S. Rekik, D. Guerchi, S. A. Selouani, and H. Hamam, "Speech steganography using wavelet and Fourier transforms," EURASIP Journal on Audio, Speech, and Music Processing 2012, no. 1, Aug 2012, pp. 1-14.

[7] F. Djebbar, B. Ayad, K. A. Meraim, and H. Hamam, "Comparative study of digital audio steganography techniques," EURASIP Journal on Audio, Speech, and Music Processing 2012, no. 1, Oct 2012, pp. 1-16.

[8] K. Gopalan and Jiajun Fu, "An Imperceptible and Robust Audio Steganography Employing Bit Modification," to be presented at the IEEE International Conference on Industrial Technology 2015, Seville, Spain, March 2015, pp. 1635-1638.

 ISBN: 978-1-61208-411-4

Stochastic Chase Decoding of BCH Codes

Lucas M. F. Harada, Daniel C. Cunha
Centro de Informática
UFPE – 50740-560
Recife – Brazil
lmfh@cin.ufpe.br, dcunha@cin.ufpe.br

Cecilio Pimentel
Department of Electronics and Systems
UFPE – 50740-550
Recife – Brazil
cecilio@ufpe.br

***Abstract*—This paper analyzes the complexity-performance trade-off of the Stochastic Chase decoding algorithm for Bose-Chaudhuri-Hocquenghem (BCH) codes over the additive white Gaussian noise (AWGN) channel. It is verified by computer simulations that this algorithm can outperform the traditional Chase algorithm with less complexity decoding.**

Keywords-Block codes; Chase algorithm; decoding complexity; additive white Gaussian noise; frame error rate; BCH code; reliability-based decoding; soft-decision decoding.

I. INTRODUCTION

Soft-decision decoding is a decoding process that utilizes the information contained in the unquantized received symbols to improve the error-correcting performance compared to hard-decision decoding. However, the better performance of the soft-decision algorithms comes at the price of higher complexity. Concerning block codes, an important class of soft-decision decoding algorithms is the reliability-based (or probability-based) decoding techniques [1]-[4].

The Chase algorithm [1] is a reliability-based decoding technique that uses a set of test patterns in attempt to find an estimation of the maximum-likelihood codeword. To generate the set of test patterns, the least reliable positions (denoted by p) of the received sequence are considered. For example, considering additive white Gaussian noise (AWGN) channel, the real values of the received sequence correspond to the reliabilities of the Chase algorithm. The higher the value of the reliability, the lower the probability that the corresponding symbol had been strongly affected by the noise. Given the p least reliable positions, the number of generated test patterns is equal to 2^p and it is a way to measure the complexity of the decoding algorithm. With respect to Reed-Solomon (RS) codes and Bose-Chaudhuri-Hocquenghem (BCH) codes, many efforts have been made to find reduced complexity Chase decoding algorithms, including for implementation of VLSI architectures [5][6].

A modification of the Chase decoding algorithm, named Stochastic Chase algorithm, was proposed in [7]. It was assumed that the test patterns are stochastically generated instead of using the least reliable positions of the received sequence. This proposal was investigated for RS codes and it was shown that this modification is a low cost solution for soft-decoding of this class of codes. However, nothing was commented about the use of the Stochastic Chase algorithm for BCH codes. With this in mind, the objective of this paper is to analyze the complexity-performance trade-off of the Stochastic Chase decoding algorithm for BCH codes and how the characteristics of the BCH code influence the performance of the decoding algorithm. For sake of simplicity, hereafter the Chase algorithm and the Stochastic Chase algorithm will be denoted, respectively, by Ch and $S-Ch$ algorithms.

The remainder of this article is structured as follows. In Section II, a modified version of the Stochastic Chase decoding algorithm is described. Section III presents numerical results. At last, conclusions are drawn in Section IV.

II. STOCHASTIC CHASE DECODING

Consider a binary linear code $C(n,k,d)$ in which n is the codeword length, k is the dimension of the code and d is the minimum Hamming distance of C. Let $\boldsymbol{v}=(v_1,v_2,\dots,v_n)$ be a codeword in C. For transmission, binary antipodal modulation and an AWGN channel are assumed. At the receiver side, the sequence of real values observed at the output of the matched filters, $\boldsymbol{r}=(r_1,r_2,\dots,r_n)$, and the binary sequence $\boldsymbol{y}=(y_1,y_2,\dots,y_n)$, obtained by hard quantization of $\boldsymbol{r}$, are used as input of the soft-decision decoding algorithm.

In Ch algorithm, the set of test patterns is given by sequences of length n which have any binary combination in the p least reliable positions. After the generation of the 2^p test patterns, they are used as input of the Berlekamp-Massey (BM) hard-decision decoder. If the decoding is successful ($\widehat{\boldsymbol{v}}$ is valid), the codeword obtained by the BM decoder is included in the set of candidate codewords $\boldsymbol{\Lambda}$. Maximum likelihood soft-decision decoding is performed for each codeword in this set.

In $S-Ch$ algorithm [7], the test pattern selection is a bit-wise stochastic experiment based on the observation of the sequence $\boldsymbol{r}$. The bit y_i of the m-th test pattern depends on the reliability r_i, which can be either represented in the probability domain as

$$p_i = P(r_i|v_i=1) = \left(1+e^{\frac{2r_i}{\sigma^2}}\right)^{-1} \quad (1)$$

where σ^2 is the AWGN power. This algorithm has three independent parameters. The variation of the threshold θ changes the number of bits that will be prevented from being inverted. Decreasing θ avoids the flipping of less reliable bits, while increasing θ prevents only the most reliable bits. The parameter β is a positive constant that must be optimized for each BCH code. The parameter τ is the total number of generated test patterns, being each one unique or

 ISBN: 978-1-61208-411-4

not. We introduce an improvement in the decoding algorithm to reduce the computational complexity compared to the original one. This consists in removing repeated test patterns. Simulation results show that for lower minimum distance codes, the number of repeated test patterns can be very high (see Section III). A summary of the proposed decoding algorithm is shown in Figure 1.

1. **for** $(1 \le i \le n)$ **do**
   if $(p_i \le 0.5 - \theta)$ **then**
   $p_i = 0$, where $0 < \theta < 0.5$
   else if $p_i \ge 0.5 + \theta$ **then**
   $p_i = 1$
   else
   $p_i = \frac{1}{1+e^{-\beta r_i}}$, where $\beta > 0$
2. **for** $(1 \le m \le \tau)$ **do**
   for $(1 \le i \le n)$ **do**
   - Generate an uniformly distributed random value: $s_i \in [0,1]$
   - Generate $y_i^m = 0$ (if $s_i \ge p_i$) or $y_i^m = 1$ (otherwise)
   - Check if $\boldsymbol{y^m}$ is not equal to previous ones. If it is duplicated, remove it.
   - Perform BM hard-decision decoding on $\boldsymbol{y^m}$ to get $\hat{\boldsymbol{v}}$. If $\hat{\boldsymbol{v}}$ is valid, insert it into the set of candidate codewords Λ.
3. Select the codeword $\boldsymbol{d}$ from Λ which has the maximum correlation with $\boldsymbol{r}$.

Figure 1. Description of the modified $S - Ch$ algorithm.

III. SIMULATION RESULTS

Computer simulations of the Ch and the $S - Ch$ algorithms were performed for BCH codes of codeword length $n = 127$ and different error-correcting capabilities ($t = 3$, 6 and 9). The three parameters of the $S - Ch$ algorithm were fixed to $\tau = 1024$, $\theta = 0.45$, and $\beta = 6$.

Table I summarizes the results obtained for the three BCH codes mentioned previously. The performance of the decoding algorithms is given by the frame error rate (FER) and the complexity metric is given by the number of BM hard-decision decoding that are performed as a step of the soft-decision decoding algorithm (Ch or $S - Ch$). For Ch algorithm, the number of BM decodings is 2^p, because every test pattern implies in one BM decoding. For $S - Ch$ algorithm, the amount of hard-decision decodings depends on the average number of distinct test patterns generated to decode each codeword (see the description of $S - Ch$ algorithm in the previous section). We denote the average number of BM decodings by N_{BM}. Both parameters (FER and N_{BM}) were obtained for both algorithms for selected values of the signal-to-noise ratio (SNR) in dB.

We observe from Table I that when $t = 6$ and SNR= 4.0 dB the $S - Ch$ with $N_{\mathrm{BM}} = 772$ outperforms the Ch with 1024 ($p = 10$) test patterns (number of BM decodings). Thus, the $S - Ch$ algorithm operating with the proposed parameters has better performance and less decoding complexity than the Ch algorithm. This trend was also observed for other values of SNR and other values of t (see, for example, the results for $t = 9$ in Table I). We also noticed that the code gain obtained by the $S - Ch$ becomes negligible for small values of t. This is observed for $t = 3$ in the table, where the values of p were selected such that 2^p is close to N_{BM}. In this case, both algorithms have comparable performance with similar complexities.

TABLE I. RESULTS OF PERFORMANCE AND COMPLEXITY OF THE CH AND S-CH ALGORITHMS APPLIED TO DIFFERENT BCH CODES OF CODEWORD LENGTH $n = 127$.

BCH (127, 106, 7) [$t = 3$]				
	Ch		$S - Ch$	
SNR	**FER**	N_{BM}	**FER**	N_{BM}
4.0	$8.37 \cdot 10^{-3}$	**512**	$6.91 \cdot 10^{-3}$	427
4.5	$1.79 \cdot 10^{-3}$	**256**	10^{-3}	237
5.0	$2.7 \cdot 10^{-4}$	**128**	$2.4 \cdot 10^{-4}$	105
BCH (127, 85, 13) [$t = 6$]				
	Ch		$S - Ch$	
SNR	**FER**	N_{BM}	**FER**	N_{BM}
4.0	$1.38 \cdot 10^{-3}$	**1024**	$5.25 \cdot 10^{-4}$	772
4.5	$1.1 \cdot 10^{-4}$	**1024**	$3.75 \cdot 10^{-5}$	607
BCH (127, 71, 19) [$t = 9$]				
	Ch		$S - Ch$	
SNR	**FER**	N_{BM}	**FER**	N_{BM}
4.0	$8 \cdot 10^{-4}$	**1024**	$2.12 \cdot 10^{-4}$	992
4.5	$5 \cdot 10^{-5}$	**1024**	$8.33 \cdot 10^{-6}$	840

IV. CONCLUSIONS

In this work, the Stochastic Chase decoding of BCH codes is investigated by a modification of the original algorithm proposed in [7]. Also, the complexity-performance trade-off of the decoding for BCH codes of codeword length $n = 127$ and different error-correcting capabilities is analyzed. Work is in progress to apply this approach to a BCH turbo decoding framework.

REFERENCES

[1] D. Chase, "A class of algorithms for decoding block codes with channel measurement information," IEEE Trans. Inf. Theory, vol. 18, n. 1, 1972, pp. 170-182.

[2] W. J. Chen, M. Fossorier, and S. Lin, "Quantization issues for soft-decision decoding of linear block codes," IEEE Trans. Commun., vol. 47, n. 6, 1999, pp. 789-795.

[3] M. Fossorier and S. Lin, "Error performance analysis for reliability-based decoding algorithms," IEEE Trans. Inf. Theory, vol. 48, n. 1, 2002, pp. 287-293.

[4] Y. Tang, S. Ling, and F. Fu, "On the reliability-order-based decoding algorithms for binary linear block codes," IEEE Trans. Inf. Theory, vol. 52, n. 1, 2006, pp. 328-336.

[5] Y. Wu, "Fast Chase decoding algorithms and architectures for Reed-Solomon codes," IEEE Trans. Inf. Theory, vol. 58, n. 1, 2012, pp. 109-129.

[6] X. Zhang, "An efficient interpolation-based Chase BCH decoder," IEEE Trans. Circuits and Systems II: Express Briefs, vol. 60, n. 4, 2013, pp. 212-216.

 ISBN: 978-1-61208-411-4

[7] C. Leroux, S. Hemati, S. Mannor, and W. J. Gross, “Stochastic Chase decoding of Reed-Solomon codes,” IEEE Commun. Lett., vol. 14, n. 9, 2010, pp. 863-865.

 ISBN: 978-1-61208-411-4

Performance Comparison of Clipping Technique with Adaptive Filters for Impulsive Noise Reduction in AWGN Environment

Sumrin Mehak Kabir, Alina Mirza, Shahzad Amin Sheikh
College of E&ME, National University of Sciences and Technology (NUST),
Islamabad, Pakistan
E-mail: sumrin.mehak75@ee.ceme.edu.pk, alinamirza2002@yahoo.com, sheikh.shahzadamin@gmail.com

***Abstract*— *Impulsive* noise is a non-Gaussian noise that appears in the communication and it may affect the information badly. In the past, methods have been investigated to mitigate this noise. In this paper, clipping based impulsive noise cancellation technique is compared with a previously existing technique comprised of Normalized Least Mean Square (NLMS) and Recursive Least Square (RLS) algorithms. The scheme is tested on binary data modulated over different types of constellation schemes. The performance of the adaptive filters is quite better than that of clipping for different modulation schemes namely Quaderature Phase Shift Keying (QPSK), 16-Quaderature Amplitude Modulation (QAM) and 32-QAM. The convergence characteristics of both methods are demonstrated by the simulation results in terms of Bit Error Rates (BER).**

Keywords- Impulsive Noise; Adaptive Filter; NLMS; RLS; QPSK; 16QAM; 32QAM; BER.

I. INTRODUCTION

Adaptive Filters with non-stationary statistical characteristics, low cost, and their ability to adapt to the unknown environment make them most suitable for the control applications and signal processing [1]. Therefore adaptive filters have been successfully used in numerous signal processing applications over the past decades.

The adaptive filter systems have general characteristics i.e., an output signal is generated by the adaptive filter and is compared with a desired signal, to generate an error. That error is then used to modify the adjustable coefficients of the filter, generally called filter tap weights, in order to minimize the error.

However practically, noise is impulsive in nature which is non-Gaussian generated by human activities [2] [3] and has more catastrophic effects in communication systems. Nowadays active area of research is to inspect the impulsive noise behavior and suggest solutions to improve the performance of systems by suppressing it. For noise cancellation, clipping technique is implemented in literature which attempt to recover the original transmitted signal [5].

In [6], the performance comparison of the adaptive filter algorithms such as the least mean square (LMS), Normalized LMS (NLMS) and Recursive least squares (RLS) were carried out to remove the noise from the audio signal. Impulsive noise has been removed using NLMS filter over different modulation schemes in [7] on the basis of step size and likelihood probabilities.

In this paper, a comparison of the already existing NLMS and RLS filters with clipping method for impulsive noise cancellation has been presented. Though clipping method is simple in terms of implementation and carrying out the parameters, it can be used for impulsive noise removal in the presence of Additive White Gaussian Noise (AWGN).

The paper is organized as: Section II briefly describes the basic principle of noise cancellation. Section III gives the review of clipping method which is followed by discussion of different adaptive filters in Section IV, supported with the simulation results in Section V. In the end, Section VI concludes the paper followed by the references.

II. IMPULSIVE NOISE GENERATOR MODEL

Impulsive noise has been generated using the model given in Figure 1 in MATLAB/Simulink [2]. It includes the zero-order hold, data source, and sign as a comparator. The output is multiplied by a random number to generate fixed or 1 unit width and variable amplitude impulses as shown in Figure 2.

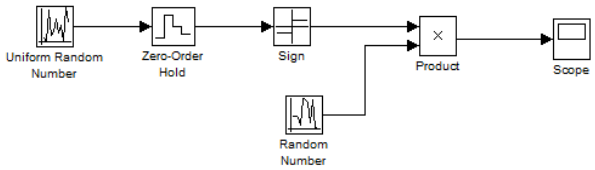


Figure 1. Impulsive noise generator model [2]

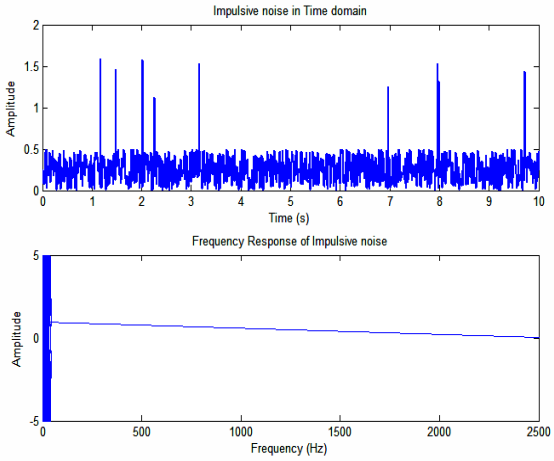


Figure 2. Impulsive noise signal and spectrum

The frequency response of the generated impulsive noise is also shown in Figure 2. The impulsive noise being a non-Gaussian noise, has a flat response having all the frequencies in equal amount.

III. NOISE CANCELLATION

Adaptive filter have an adaptation algorithm that monitors the environment and vary the filter transfer function accordingly. Each adaptive filter depends on the error signal computed from the adaptive filter to update its filter taps.

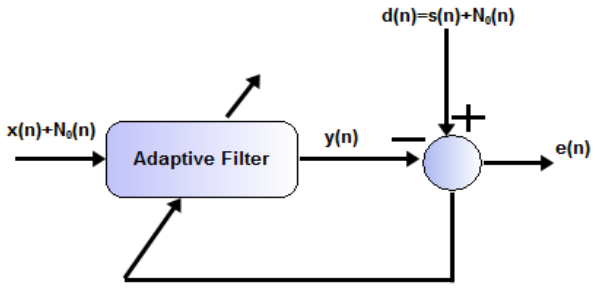


Figure 3. Block Diagram of noise cancellation in adaptive filters

The output of the adaptive filter is compared with the desired signal *d(n)*.The desired signal is a random signal corrupted with another noise source or system noise, as shown in Figure 3. The coefficients of adaptive filters adapt recursively to make the error signal *e(n)* to be minimum so that the output signal *y(n)* approaches to be equivalent to the desired signal having minimum error.

The interference cancellation application of the adaptive filters is employed to cancel the noise from the signal. For that, the adaptive filter updates recursively in order to remove the noise from the input signal using the noise in the desired signal by subtracting it from the output signal.

IV. CLIPPING

In practical applications, clipping method is usually used for impulsive noise mitigation due to its simplicity. A clipping algorithm is employed at the receiver end of the AWGN channel, where we presume that impulsive noise is being added by the channel itself during the communication [5]. It is to reckon that clipping method only changes the amplitude of the data without changing the other parameters such as phase.

$$y_k = \begin{cases} r_k & , if\ r_k < T_c \\ T_c\, e^{jarg(r_k)} & , if\ r_k \geq T_c \end{cases} \quad (1)$$

T_c is the Clipping Threshold that can be set according to the maximum value of data if known. And r_k is a sample of the signal to be clipped over T_c. The $arg(r_k)$ is used due to the possibility of the existence of complex valued samples in the signal.

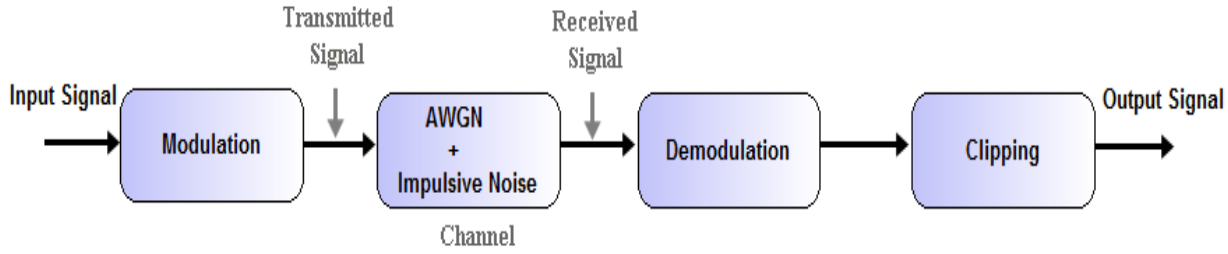


Figure 4. Block diagram of Clipping method

In this method the amplitude of the received data is clipped or limited by the threshold and no other change occurs to the signal/data that has been received. Figure 4 shows the simple model carried for the clipping method.

V. ADAPTIVE FILTERS

There are many adaptive algorithms used for noise removal. The brief summaries of adaptive algorithms which are used in this research are as follows.

A. NLMS

The Normalized Least Mean Square (NLMS) is one of the variants of LMS algorithm whose convergence ate is faster than LMS. The drawback of LMS algorithm is its sensitivity.

To avoid gradient noise amplification problem NLMS algorithm is used in which the tap weight *w(n)* at

 ISBN: 978-1-61208-411-4

n+1 is normalized with the Euclidean norm of the square of the input to the filter so that the convergence is stable. The filter tap weights are updated using following equation during the recursive procedure:

$$w(n+1) = w(n) + \frac{\mu e(n) x(n)}{\epsilon + \|x(n)^2\|} \quad (2)$$

Where ϵ is a small number added for algorithm stability, μ is the step size and $e(n)$ is error signal.

B. RLS

The Recursive least squares (RLS) adaptive filter belongs to the least square family of the adaptive filters. It tends to minimize a linear least cost function related to the input signal by finding the coefficients recursively. Also the input signal for RLS is deterministic. The convergence rate of RLS is far higher than many other adaptive algorithms. However, it costs in higher computational complexity. The weights of the filter are updated by these equations:

$$w(n+1) = w(n) + k(n)x(n) \quad (3)$$

$$k(n) = \frac{\lambda^{-1}\Phi^{-1}(n-1)x(n)}{1+\lambda^{-1}x^T(n)\Phi^{-1}(n-1)x(n)} \quad (4)$$

$$\Phi^{-1}(n) = \lambda^{-1}\Phi^{-1}(n-1) - \lambda^{-1}k(n)x^T(n)\Phi^{-1}(n-1) \quad (5)$$

Where λ is the forgetting factor. Φ^{-1} is the cross correlation matrix.

VI. Simulation Results

In the first part of simulations, the binary data perturbed by impulsive noise in Figure 5, is recovered using clipping method. This is shown by initially generating impulsive noise by following steps mentioned in [2] and depicted in Figure 2.

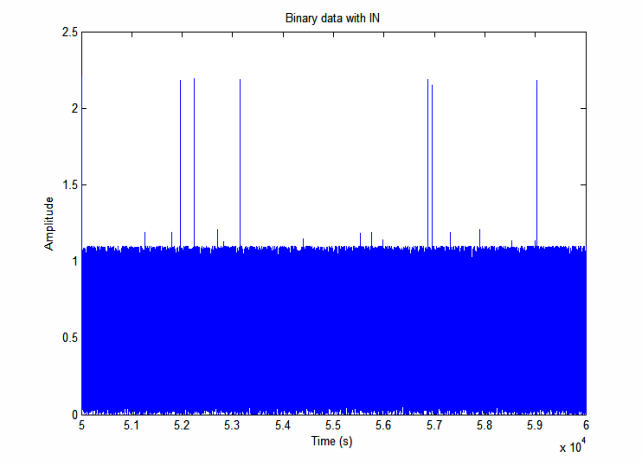


Figure 5. Binary input signal with Impulsive noise

The binary data is generated randomly, comprised of 100,000 randomly generated bits, in MATLAB. This binary data is added with the impulsive noise after modulation as shown in Figure 5 and is then transmitted over an AWGN channel, pretending that impulsive noise has been added during the communication.

For clipping method, the received binary data is clipped over a threshold of amplitude 1 after demodulation. It can be put in the way that the clipping method can clip or limit the data above the threshold but it cannot remove the noise that lie below the threshold. The noise below the threshold is usually caused by additive white Gaussian noise from the channel.

In the second part of the simulations, in adaptive algorithms the demodulated data is filtered using the adaptive filters (NLMS & RLS). The basic scenario is that the output $y(n)$ of the filter is compared with the desired signal *s(n)*. Their difference produces the error signal *e(n)* and the filter updates its weights recursively using the adaptive weight update equations (2) and (5), such that the error signal is minimized.

$$\begin{aligned} e(n) &= d(n) - y(n) \\ &= s(n) + n_2(n) - n_2(n) = s(n) \end{aligned} \quad (6)$$

In an optimum sense, the system output signal should contain the original signal as in (6). For these simulations the length of the two adaptive filters is fixed to 32. The step size parameter μ for NLMS Algorithm is chosen to be 0.1 and forgetting factor λ for RLS is 0.98.

Figure 6 represents the bit error rate (BER) plot of clipping method, NLMS algorithm and RLS algorithm for QPSK modulation. The performance of IN reduction methods is compared through the BER plots of the received signal that does not contain impulsive noise and is considered to show the minimum number of errors occurred in the signal, due to the channel only.

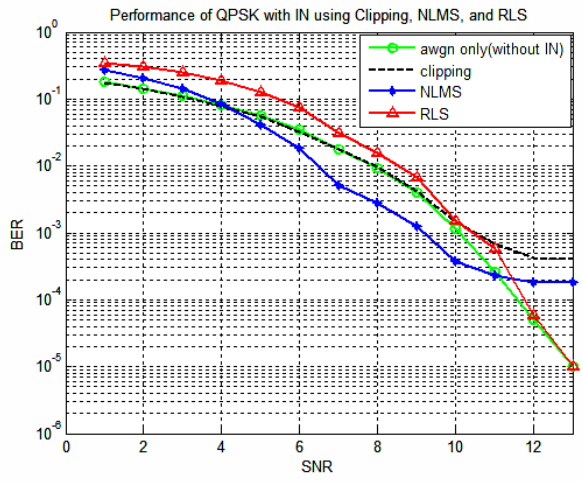


Figure 6. Comparison of BER (dB) over QPSK modulation

It can be seen that among the three algorithms, the performance of RLS is quite better than the rest of the methods for an SNR of up to 20 dB.

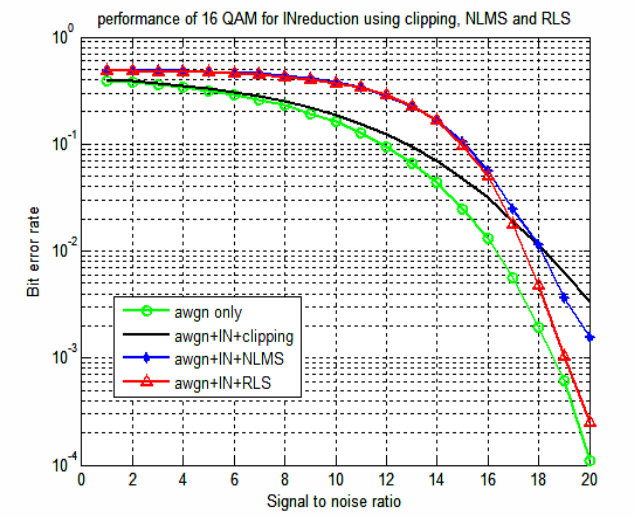


Figure 7. Comparison of BER (dB) over 16-QAM modulation

For 16-QAM modulations or square constellation, the results show that among the performances of the three methods, RLS is better as shown in Figure 7. Similarly, Figure 8 shows the BER comparison for 32-QAM modulation that is a rectangular QAM constellation and the results are hereby proven that RLS depicts far better performance in comparison with the clipping method.

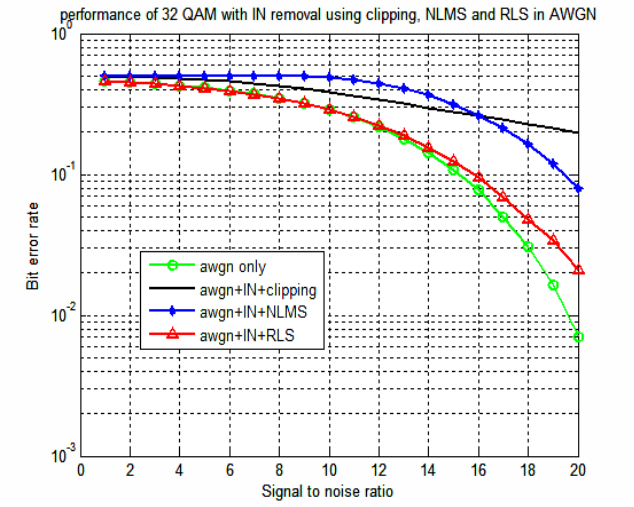


Figure 8. Comparison of BER (dB) over 32-QAM modulation

The performance metric bite error rate in decibel indicates that RLS filter has lowest BER compared to other used algorithms. The BER of RLS filter is close to the performance of minimum possible errors for the removal f impulsive noise. Whereas the BERs of NLMS and clipping method are much greater than the RLS BER and the AWGN only performance that has no impulsive noise.

VII. Conclusion

In this paper, adaptive noise cancellation technique based on Normalized Least Mean Square (NLMS) and Recursive Least Square (RLS) algorithm, are compared with another previously existing technique named as clipping. Due to recursive parameters, the adaptive filters require the reference signal and exhibit better impulsive noise cancellation when compared with the clipping technique. The conducted comparison guarantees that adaptive RLS filter is an efficient noise canceller for different modulation schemes such as QPSK, 16 QAM (square constellation) and 32 QAM (rectangular constellation). It ensured the better performance of RLS in terms of convergence speed and lower BER and is verified by the simulation results.

References

[1] B. F. Boroujeny, Adaptive Filters: Theory and Applications, John Wiley, BaffinsLane, Chichester, 1998.

[2] S. R. Al-Araji, M. A. Al-Qutayri, and M. S. Al-Tenaiji, “Impulsive noise reduction using Auto-Gating technique,” GCC Conference and Exhibition (GCC), 2011 IEEE, 19-22 Feb, 2011, pp.104-107.

[3] S. V. Vaseghi, “Advanced Digital signal processing and noise reduction,” John Wiley & Sons Ltd. Fourth edition, 2000. ISBN : 978-0-470-75406-1

[4] S. Haykin, “Adaptive filter theory,” 3rd Edition, Prentice Hall, 1996.

[5] S. Al-Mawali, and Z. H Hussain, “Adaptive-threshold clipping for impulsive noise reduction in OFDM-based power line Communications,” International Conference on Advanced Technologies for Communications, 12-14 Oct, 2009, pp.43-48.

[6] R. K. Thenua, and S. K. Agarwal, “Simulation And Performance Analysis Of Adaptive Filter In Noise Cancellation,” International Journal of Engineering Science and Technology, vol. 2(9), 2010, pp. 4373-4378.

[7] S. A. Jimaa, S. R. Al-Araji, A. Al-Kaabi, and T. Shimamura, “Impulsive Noise Reduction using Adaptive Receiver Structure Technique,” IEEE 11th International Conference on Signal Processing (ICSP), Beijing, China, vol. 1, 21-25 October, 2012, pp. 119-122.

 ISBN: 978-1-61208-411-4

Mitigating Distributed Denial-of-Service Attacks in Named Data Networking

Vassilios G. Vassilakis
Institute for Communication Systems
University of Surrey
Guildford, United Kingdom
e-mail: v.vasilakis@surrey.ac.uk

Bashar A. Alohali
School of Computing and Mathematical Sciences
Liverpool John Moors University
Liverpool, United Kingdom
e-mail: b.a.alohali@2012.ljmu.ac.uk

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

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

***Abstract*—Named Data Networking (NDN) is a novel networking approach that aims at overcoming some of the limitations of the current Internet. In particular, NDN aims at providing better privacy and security by focusing on the data items themselves rather than on the location of data. This is achieved by using soft states at the routers, which record the requests/interests for data from users in the Pending Interest Table (PIT). However, this new networking concept opens up avenues for launching Distributed Denial-of-Service (DDoS) attacks on PITs. That is, an attacker may flood the network with a large number of Interest packets that would overflow the PITs at the routers, thus preventing legitimate users from receiving the requested data. This type of DDoS attack is known as the Interest Flooding Attack (IFA) and, if not adequately dealt with, may severely disrupt the normal operation of an NDN system. In this paper, we first show that the basic NDN mechanism is vulnerable to IFA even when the attacker has very limited resources. Next, we propose a mitigation technique that allows routers to quickly identify and block such DDoS attempts, by detecting anomalous user behaviour. We also introduce an additional security layer by using public-key based router authentication. We evaluate our proposed scheme by means of computer simulations and show that a sufficient level of security can be achieved with little processing and storage overhead.**

Keywords–Named Data Networking; Distributed Denial of Service; Interest Flooding Attack.

I. INTRODUCTION

As it has been observed by numerous studies, today the Internet is mainly used for data dissemination to interested users, rather than for connecting hosts. The user is interested in data itself, while the *location* of data is usually of minor importance. However, the Internet was originally designed and has evolved according to the host-centric communication paradigm. Recent studies have shown that the poor performance of the traditional Internet, in the areas of security, efficient content dissemination, etc., lies in its host-centric nature [1].

Information Centric Networking (ICN) [2] is a new effort that aims at eliminating the traditional Internet's limitations. Named Data Networking (NDN) [3] is one of the proposed ICN approaches. Data dissemination in NDN is achieved by using soft states at the routers, which record the interests for data from users in the Pending Interest Table (PIT) [4]. When the requested data is received by the router and forwarded to the user, the corresponding PIT entry is removed. In contrast to the current Internet, where the security measures have been added after its conception, NDN's key target is to embed security and privacy features at the very early design stages. In particular, for protecting user privacy, no source address is carried in the packets [5]. The routers record in PIT the incoming interface for the Interest packet and use it to forward the data to the user. NDN also inherently provides protection against unsolicited data by adopting the receiver-driven data retrieval model [6].

However, in spite of the aforementioned security advantages of NDN, new types of distributed denial-of-service (DDoS) attacks are possible [7]. In particular, one of the most challenging is the DDoS attack on PITs, where an attacker floods the network with a large number of bogus Interest packets. Each such packet causes the router to create and maintain an entry in its PIT, thus wasting router's storage resources and even creating the possibility of PIT overflows. This type of attack is known as the Interest Flooding Attack (IFA) [8] and, if not adequately dealt with, may severely disrupt the normal operation of an NDN system.

In this paper, we first show that the basic NDN mechanism is vulnerable to IFA even when the attacker has very limited resources. Next, we propose an IFA mitigation technique at the router, that detects anomalous user behaviour and also notifies other routers. To secure from bogus notifications by malicious/compromised routers, we propose an authentication scheme that is based on public-key cryptography.

This paper is organized as follows. In Section II, we briefly describe the considered NDN architecture and introduce the necessary notations. In Section III, we define the IFA and briefly discuss other types of DDoS attacks in NDN. In Section IV, we present our IFA mitigation scheme. In Section V, we present our router authentication method. In Section VI, we study the performance of the proposed approach by means of computer simulations. In Section VII, we present the related work on DDoS attacks and countermeasures in NDN and in other ICN mechanisms. We conclude and discuss our future work in Section VIII.

 ISBN: 978-1-61208-411-4

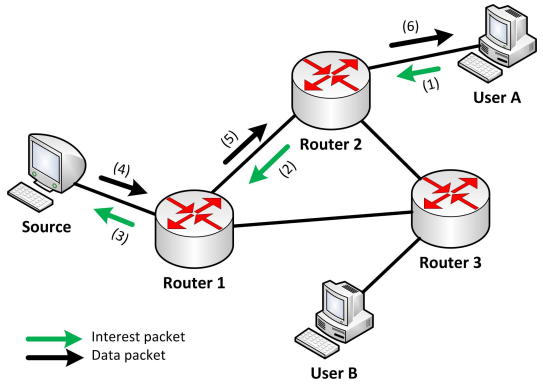


Figure 1. Basic NDN communication model.

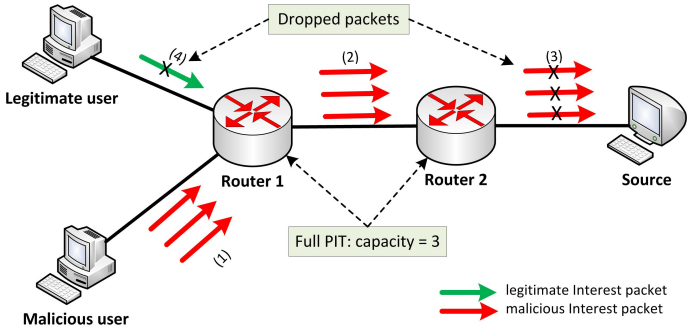


Figure 2. Interest Flooding Attack on NDN.

II. NDN Architecture and Basic Concepts

In this section, we briefly describe the NDN architecture and its basic concepts [3]. Contrary to the traditional host-centric network architectures (e.g., the Internet), the basic abstraction in NDN is the *named content*. Content sources advertise/publish their available content items in the network by issuing the *Publication packets*, which usually include prefix-based content name or some other form of content identifier (ID) [9]. Each router, upon receiving such packets, records the incoming interface and the content name in its Forwarding Information Base (FIB) [10].

ICN also natively supports quality-of-service [11] and content caching. The caching approach could be either opportunistic and transparent to the content resolution function (i.e., not recorded in the FIB) or proactive and similar to the Content Delivery Network (CDN) approach [12]. The latter approach essentially transforms caches into alternative content sources and enables joint optimization of forwarding and caching functions [13].

Users interested in receiving a particular content item, issue the *Interest packets*. These packets are then forwarded via a sequence of routers towards the content source or cache, according to the FIB entries for the requested content. In case that multiple sources (and/or caches) are holding the same content, some kind of mediated topology management function could be used to select the best source [14] or even to enable multi-chunk content delivery [15].

When the source/cache receives the Interest packet it replies with the *Data packet*, which contains the requested content. The Data packet is forwarded via the reverse path towards the user. A simple illustrative example is shown in Figure 1. This is the so-called pull-based communication model and it ensures that the user receives only explicitly requested content.

In order for the routers to be able to deliver Data packets to the users, each router is equipped with a PIT [4]. The latter contains entries for all "not yet satisfied" Interest packets and their incoming interface. Note that this communication scheme does not require any user address (e.g., IP address) carried in the Interest and Data packets. These packets are required to carry only the content name of some other kind of content ID.

III. DDoS in NDN

A. Interest Flooding Attack

Malicious/compromised users may exploit the PIT-based forwarding mechanism of NDN to launch the IFA, which is considered as one of the most serious types of DDoS attacks on NDN [16]. According to IFA, the malicious user (or a group of users) will issue a large number of bogus Interest packets. Each router, upon receiving each of these packets, will create an entry in its PIT and will forward the packet to the next-hop node (router or content source). According to the NDN rules, an entry is removed from the PIT in the following two cases:

- Entry expired (e.g., a typical expiry time is 1s [17]).
- Router received the corresponding Data packet before the entry expiration.

According to the above, the best attacking strategy is to issue Interest packets for non-existent content. In this case, the bogus entries will stay in the PIT as much as possible. The goal of the attacker is to quickly fill in the PIT and to keep it full, so that the Interest packets originated from legitimate users will eventually be dropped.

In Figure 2, we illustrate a simple example of IFA in NDN. Assume that the PIT capacity in each router is 3 entries. The attacker's strategy is to send 3 bogus Interest packets for (different) non-existent content. These packets will fill in the PITs of both routers. The source will drop these packets, since they request non-existent content. However, the corresponding entries will stay in the PITs until they expire. After the expiration, the attacker will issue 3 new Interest packets, aiming at keeping the PITs always full. This way, some, or even all, Interest packets of legitimate users will be dropped. Later, in Section VI, we evaluate the packet dropping probability of legitimate users and show that it could be very high even when the attackers employ limited resources.

B. Other DDoS Attacks

In this subsection, we briefly discuss other possible types of DDoS attacks in NDN, which, however, are out of the scope of this paper.

1) cache poisoning/pollution attack: The attacker is trying to reduce the cache efficiency by filling in the cache with non-popular or even fake content. This can be done by repeatedly requesting the same unpopular content. This aims at increasing the cache misses and forcing the Interest packets to reach the content source. This type of attack is not easy to mitigate,

 ISBN: 978-1-61208-411-4

because the malicious user may appear as a legitimate one for a very long time.

2) mobile interest flooding attack: The attacker may periodically visit different routers and issue bogus Interest packets. This attack is harder to detect and mitigate than the classic IFA. The reason is that retransmission of Interest packets in case of mobility is a normal procedure in NDN. So, to detect a mobile attacker, a complex scheme that involves a large number of cooperating routers would be required.

3) attack on forwarding mechanism: A potentially compromised router may severely degrade the performance of the network by re-directing the Interest packets in wrong direction. In case of cooperating attackers, this could even be exploited for creating forwarding loops in the network.

IV. IFA Mitigation Mechanism

In this section, we describe our proposed mitigation mechanism for IFA in NDN. The aim of this approach is to quickly detect anomalous user behaviour and to restrict, or even block, such user at an early stage of the attack. We distinguish two types of routers:

- *Edge routers*: directly connected to one or more users.
- *Core routers*: directly connected only to other routers or sources.

The edge routers will provide an additional security layer by detecting any anomalous user behaviour and will notify other routers if such an event takes place. The latter is done by sending the *attack notification packets*, that contain the user ID. Core routers will be involved in forwarding the attack notification packets to other routers, but will not themselves contribute in the attack detection process.

Our mitigation mechanism comprises three phases:

- *Attack detection phase*: the edge router detects anomalous user behaviour and identifies the user either as suspicious or as an attacker.
- *Rate reduction and blocking phase*: the edge router reduces the data rate of suspicious users and blocks the attackers.
- *Attack notification phase*: the edge router notifies other edge routers about the detected attack.

In the following, we provide more details about these three phases.

A. Attack detection phase

The set of all users in the network is denoted by U. During the detection phase, the edge router keeps statistics about the expired PIT entries per each user $u \in U$. Two thresholds are used to classify users into: *legitimate*, *suspicious* (possible attackers), and *malicious* (attackers). If the number of expired PIT entries per time unit, $N_{exp}(u)$, of a user u is below the low threshold, T_{low}, user u is considered legitimate. If $N_{exp}(u)$ is above T_{low} but below the high threshold, T_{high}, user u is considered suspicious. Finally, if $N_{exp}(u) > T_{high}$, user u is considered malicious. The sets of legitimate, suspicious, and malicious users are denoted by L, S, and M, respectively.

B. Rate reduction and blocking phase

During this phase, any user that has been classified as malicious, will be blocked, whereas the suspicious users will receive reduced data rate. In particular, the rate adaptation is performed as follows:

$$R_{new}(u) = \begin{cases} R_{old}(u), & \text{if } u \in L \\ \frac{aR_{old}(u)}{T_{high}-T_{low}}, & \text{if } u \in S \\ 0, & \text{if } u \in M \end{cases} \quad (1)$$

where $R_{old}(u)$ and $R_{new}(u)$, are the old and new data rate of user u, respectively; a is some optimization parameter, such that $a + T_{low} < T_{high}$.

C. Attack notification phase

If an edge router detects an ongoing attack, after blocking this user, it will notify other routers about the identity of the malicious user, by sending the attack notification packet. This is done to prevent the Mobile Interest Flooding Attack (MIFA) [7], where a mobile user periodically visits different routers and floods them with Interest packets. In this context, the notion of router is extended and refers to any data-forwarding network element, such as a WiFi Access Point (AP) or a Base Station (BS) in a cellular network.

V. Router Authentication Method

We consider a scenario where edge routers may be deployed by home users or other non-trusted parties. That will most likely be the case in future fifth generation (5G) cellular networks [18], in future generation Internet [19], and in smart grid networks [20]. Also, edge routers may join and leave the network or change their location (e.g., vehicular communications). This introduces new security threats and a good authentication method is needed.

In this section, we propose a public-key based router authentication method (similar to [21]) to protect against bogus *attack notification packets* that could be sent by potentially compromised routers. Our method makes the reasonable assumption that there will be at least one trusted network entity that can act as Certificate Authority (CA). We consider the following two cases:

- *Direct authentication*: Performed when the new router has a direct connection with CA.
- *Indirect authentication*: Performed when the new router has no direct connection with CA. In that case, the authentication of a new router is facilitated by another, already authenticated router, referred to as *mediator*.

A. Direct authentication

Initially, the new router will send the *authentication request* message to CA (see also Figure 3).

This message is encrypted using CA's public key, PK-CA, and includes the following information:

a) identity number of the new router, R-ID,
b) timestamp, TS,
c) symmetric key of new router, SK-R.

CA responds with the *authentication response* message.

 ISBN: 978-1-61208-411-4

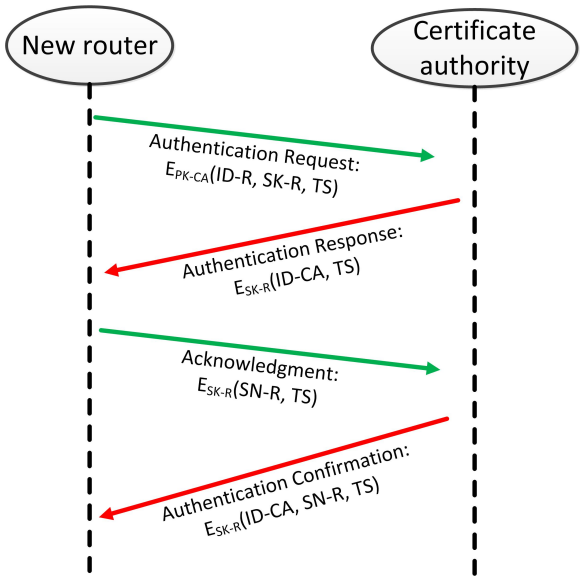


Figure 3. Direct router authentication.

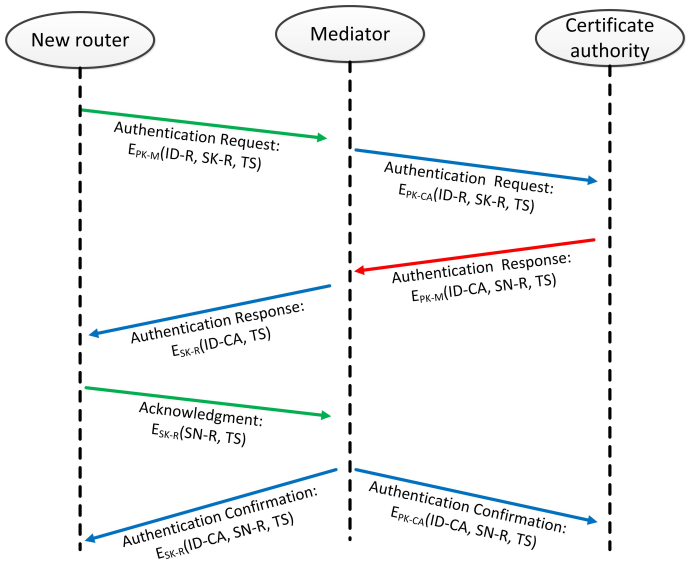


Figure 4. Indirect router authentication.

This message is encrypted using the received SK-R and includes the following information:

a) identity number of CA, ID-CA,
b) timestamp, TS.

Next, the router sends the *acknowledgement* message, which also contains routers serial number, SN-R, so that CA may validate this router as a legitimate one (we assume that CA holds a list of all valid SN-Rs).

This message is encrypted using the SK-R and includes the following information:

a) serial number of new router, SN-R,
b) timestamp, TS.

Finally, CA responds with the *authentication confirmation* message. This message is encrypted using the received SK-R and includes the following information:

a) identity number of CA, ID-CA,
b) timestamp, TS,
c) serial number of new router, SN-R.

B. Indirect authentication

Initially, the new router will send the *authentication request* message to the mediator (see Figure 4).

This message is encrypted using mediator's public key, PK-M, and includes the following information:

a) identity number of new router, ID-R,
b) timestamp, TS,
c) symmetric key of new router, SK-R.

The mediator will decrypt the message using its private key. Next, it will encrypt the message using CA's public key, PK-CA, and will send it to CA.

CA responds to mediator with the *authentication response* message. This message is encrypted using mediator's public key, PK-M, and includes the following information:

a) identity number of CA, ID-CA,
b) timestamp, TS,
c) serial number of new router, SN-R.

The mediator will decrypt this message using its private key. Next, it will remove the SN-R from the message, will store it and will re-encrypt the remaining message using the previously received SK-R and will send the message to the new router. The serial number, SN-R, will be used later by the mediator to verify that the new router is legitimate.

Next, new router sends to mediator the *acknowledgment* message that contains its serial number, SN-R, to be used for validation by mediator.

This message is encrypted using SK-R and includes the following information:

a) serial number of new router, SN-R,
b) timestamp, TS.

Finally, if the authentication is successful, the mediator responds to both new router and CA with the *authentication confirmation* message. This message is encrypted for new router using SK-R and for CA using PK-CA, and includes the following information:

a) identity number of CA, ID-CA,
b) timestamp, TS,
c) serial number of new router, SN-R.

VI. Performance Evaluation

In this section, we evaluate the DDoS vulnerability of the basic NDN mechanism [3] described in Section II. To this end, we have developed an NS-3 based simulator for NDN and implemented our proposed attack mitigation scheme for a randomly generated network topology. We have simulated 1,000 legitimate users generating traffic at a rate of 20 packets/sec. We have also simulated a DDoS attack scenario on PIT, where 50 attackers generate bogus Interest packets at a rate of 1,000 packets/sec. The packet size is selected to be 1KB, so that the attacking capability of each attacker is 1Mbps.

 ISBN: 978-1-61208-411-4

The duration of our simulation is 50s. The attack starts at $t = 5$s and lasts until $t = 40$s. In Figure 5, we show the required size of the PIT of the edge router, as it grows over time due to the launched attack. The results are shown for three different PIT entry expiration times, $t_{exp} = 200$ms, 500ms, and 1s. As discussed in Section III, $t_{exp} = 1$s is the currently adopted value in NDN. The value $t_{exp} = 200$ms is certainly too optimistic (in terms of PIT size requirements) and is shown only for comparison, as the best case scenario for the victim router. With such small t_{exp}, a large number of legitimate requests will not be satisfied, since the corresponding PIT entry of each request will expire before the content arrives. In Figure 5, we observe that during the first 5 seconds, when only legitimate users are active, the required PIT size is relatively small ($\approx$ 20MB). However, shortly after the attack starts, the PIT size increases rapidly and almost reaches 2GB by $t = 40$s, for $t_{exp} = 1$s. These results show that it is relatively easy even for attacking nodes of limited capabilities to quickly occupy large amounts of router's storage and processing resources.

In the second phase, to show the negative impact of the IFA, we evaluate the packet dropping probability of legitimate users due to PIT overflow. We consider the PIT capacity equal to 1GB and use $t_{exp} = 500$ms (with $t_{exp} = 1$s the negative impact on victims would be even worse). The rest of the simulation parameters remain the same as described in the previous paragraph. In Figure 6, we present the results for the basic NDN mechanism and for our threshold-based mitigation scheme of Section IV. In the latter, the attack detection thresholds are chosen to be $T_{high} = 0.5$ and $T_{low} = 0.25$. We consider two cases with different optimization parameters: $a = 1/8$ and $a = 1/12$. When $a = 1/8$, the mitigation scheme reduces the data rate to 50%. That is, $R_{new}(u) = \frac{1}{8}\frac{R_{old}(u)}{0.5-0.25} = 0.5R_{old}(u)$. Similarly, when $a = 1/12$, the mitigation scheme reduces the data rate to 33%. In Figure 6, we observe that when no mitigation scheme is used (i.e., basic NDN mechanism is assumed), 80% of the Interest packets of legitimate users are dropped between around $t = 25$s and $t = 40$s. This is due to the fact that by $t = 25$s the PIT size has reached its capacity of 1GB. When the mitigation scheme is used, the worst-case dropping probability is reduced to 60% and 40%, for $a = 1/8$ and $a = 1/12$, respectively. Also, the negative impact of the PIT overflow is time shifted (by 5s when $a = 1/8$ and by 8s when $a = 1/12$). We have also tried $a = 1/16$ which results in 25% data rate reduction for malicious users and completely eliminates the packet dropping of legitimate users.

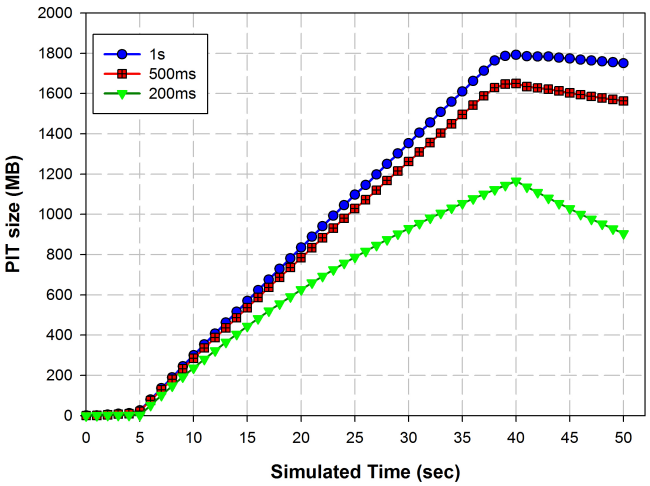


Figure 5. PIT size for different PIT entry expiration times. The attack window is 5-40s.

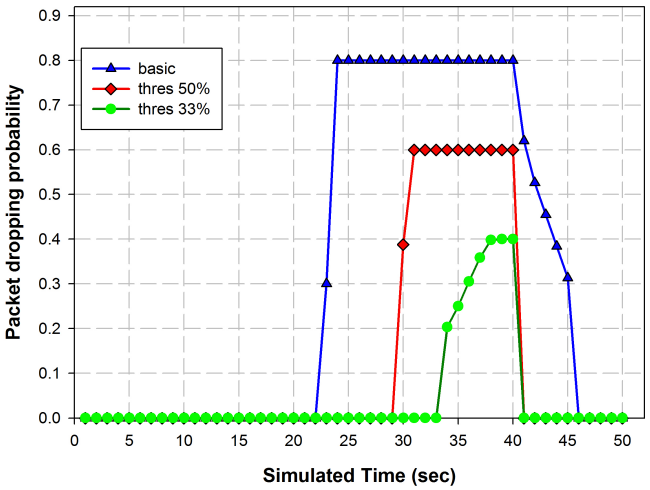


Figure 6. Packet dropping probability of legitimate users for the basic NDN and two threshold-based mitigation schemes.

VII. Related Work

A number of works study DDoS attacks in NDN. In [16], various types of attacks and possible countermeasures are discussed. It is argued that the most difficult to mitigate are the IFA and the cache poisoning attacks. However, no evaluation or assessment is presented. In [22], the *token bucket* method is proposed to mitigate the IFA. According to this method, the routers are restricted in forwarding Interest packets based on the load of their outgoing interfaces. To enable such behaviour, the routers need to keep track of the requested data volume from each interface. The proposed approach has been evaluated using the ndnSIM simulator [23] and shows satisfactory performance in cases of moderate attacking capability. In [24], to alleviate the negative impact of the IFA on PIT, the Disabling PIT Exhaustion (DPE) mechanism is proposed. Interest packets are dynamically diverted out of the PIT if their prefixes are detected as malicious. However, this introduces extra overhead on routers for marking packets and for maintaining a malicious list to be used when the PIT is exhausted. In [25], some attack scenarios similar to [16] are presented and a DDoS attack mitigation technique based on Interest traceback is proposed. According to this technique, when the content source receives a bogus Interest packet, it will send the traceback message on the reverse path to notify the involved routers. The proposed solution is effective when the round-trip time (RTT) is relatively small and if the attacker does not change location (i.e., the solution is not effective against the MIFA). In [17], the Poseidon scheme is proposed, which focuses on early detection of IFA and its subsequent mitigation. The IFA mitigation is performed by reducing the data rate of the incoming interfaces. However, this approach may also degrade the performance of legitimate users and needs further enhancements in terms of user differentiation.

 ISBN: 978-1-61208-411-4

Other ICN approaches, e.g., such as PURSUIT [26], are also vulnerable to DDoS attacks but not to IFA. Contrary to NDN, PURSUIT adopts stateless data forwarding mechanism [27] and, therefore, does not suffer from attacks (such as IFA) that target soft states. In PURSUIT, the content delivery path is included in the packet header, in the form of a Bloom filter (BF). The latter, although provides time- and space-efficient path representation, suffers from the effect of false positives during the packet forwarding [28]. False positives can be exploited to launch DDoS on both network infrastructure and end users. Some of the proposed solutions include advanced encryption [29] and authentication techniques [30], in-packet BF size optimization [31], and false positives reduction [32].

VIII. Conclusion and Future Work

In this paper, we evaluate the DDoS vulnerability of NDN through simulations. In particular, we consider the IFA, where malicious users are trying to saturate the PIT and to disrupt the normal network operation. We show, that, if no adequate countermeasures are taken, it is relatively easy to cause PIT overflow and to achieve an 80% packet dropping rate. Next, we propose an IFA mitigation scheme that is based on anomalous user behaviour detection. If a user exceeds a predefined threshold it is forced to reduce its data rate or may even be blocked. This scheme is shown to be able to significantly reduce the PIT size, in terms of bogus entries, and to improve the QoE of legitimate users, in terms of packet dropping probability. Finally, we present a public-key based authentication scheme to protect the network against malicious notification messages from compromised routers. In our future work, we are planning to develop a stochastic model for PIT and to analytically determine the PIT size and the packet dropping probability.

References

[1] A. Feldmann, "Internet clean-slate design: What and why?," ACM SIGCOMM Computer Commun. Review, vol. 37, 2007, pp. 59-64.

[2] G. Xylomenos, et al., "A survey of information-centric networking research," IEEE Communications Surveys & Tutorials, vol. 16, no. 2, 2014, pp. 1024-1049.

[3] V. Jacobson, D. Smetters, J. Thornton, M. Plass, N. Briggs, and R. Braynard, "Networking named content," Proc. CoNEXT, Rome, Italy, Dec. 2009, pp. 11-18.

[4] H. Yuan and P. Crowley, "Scalable pending interest table design: From principles to practice," Proc. IEEE INFOCOM, 2014, pp. 2049-2057.

[5] S. Arianfar, T. Koponen, B. Raghavan, and S. Shenker, "On preserving privacy in content-oriented networks," Proc. ACM SIGCOMM Workshop on Information-Centric Networking, 2011, pp. 19-24.

[6] C. Ghali, G. Tsudik, and E. Uzun, "Network-layer trust in named-data networking," ACM SIGCOMM Computer Communication Review, vol. 44, no. 5, 2014, pp. 12-19.

[7] M. Wählisch, T. Schmidt, and M. Vahlenkamp, "Lessons from the past: Why data-driven states harm future information-centric networking," IFIP Networking Conference, 2013, pp. 1-9.

[8] S. Choi, K. Kim, S. Kim, and B. Roh, "Threat of DoS by interest flooding attack in content-centric networking," Proc. IEEE International Conference on Information Networking (ICOIN), 2013, pp. 315-319.

[9] F. Li, F. Chen, J. Wu, and H. Xie, "Longest prefix lookup in named data networking: How fast can it be?," Proc. 9th IEEE NAS, 2014, pp. 186-190.

[10] C. Yi, A. Afanasyev, I. Moiseenko, L. Wang, B. Zhang, and L. Zhang, "A case for stateful forwarding plane," Computer Communications, vol. 36, no. 7, 2013, pp. 779-791.

[11] M. F. Al-Naday, A. Bontozoglou, V. G. Vassilakis, and M. J. Reed, "Quality of service in an information-centric network," Proc. IEEE GLOBECOM, Austin, USA, December 2014, pp. 1861-1866.

[12] A. Vakali and G. Pallis, "Content delivery networks: Status and trends," IEEE Internet Computing, vol. 7, no. 6, 2003, pp. 68-74.

[13] V. G. Vassilakis, et al., "A cache-aware routing scheme for information-centric networks," Proc. 9th IEEE/IET CSNDSP, 2014, pp. 721-726.

[14] B. A. Alzahrani, M. J. Reed, J. Riihijärvi, and V. G. Vassilakis, "Scalability of information centric networking using mediated topology management," Journal of Network and Computer Applications, vol. 50, April 2015, pp. 126-133.

[15] L. Wang, S. Bayhan, and J. Kangasharju, "Optimal chunking and partial caching in information-centric networks," Computer Communications, vol. 61, May 2015, pp. 48-57.

[16] P. Gasti, G. Tsudik, E. Uzun, and L. Zhang, "DoS and DDoS in named data networking," Proc. 22nd IEEE International Conference on Computer Communications and Networks (ICCCN), 2013, pp. 1-7.

[17] A. Compagno, M. Conti, P. Gasti, and G. Tsudik, "Poseidon: Mitigating interest flooding DDoS attacks in named data networking," Proc. 38th IEEE Conf. on Local Computer Networks (LCN), 2013, pp. 630-638.

[18] P. Demestichas, et al., "5G on the horizon: Key challenges for the radio-access network," IEEE Vehicular Technology Magazine, vol. 8, no. 3, 2013, pp. 47-53.

[19] J. Pan, S. Paul, and R. Jain, "A survey of the research on future Internet architectures," IEEE Commun. Mag., vol. 49, no. 7, 2011, pp. 26-36.

[20] J. S. Vardakas, N. Zorba, and C. V. Verikoukis, "Performance evaluation of power demand scheduling scenarios in a smart grid environment," Applied Energy, vol. 142, 2015, pp. 164-178.

[21] B. A. Alohali and V. G. Vassilakis, "Secure and energy-efficient multicast routing in smart grids," Proc. 10th IEEE Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), Singapore, April 2015, pp. 12-17.

[22] A. Afanasyev, P. Mahadevan, I. Moiseenko, E. Uzun, and L. Zhang, "Interest flooding attack and countermeasures in named data networking," Proc. IFIP Networking Conference, 2013, pp. 1-9.

[23] A. Afanasyev, I. Moiseenko, and L. Zhang, "ndnSIM: NDN simulator for NS-3," NDN Project, Tech. Rep. NDN-0005, 2012.

[24] K. Wang, H. Zhou, Y. Qin, J. Chen, and H. Zhang, "Decoupling malicious interests from pending interest table to mitigate interest flooding attacks," Proc. IEEE Globecom Workshops (GC Wkshps), 2013, pp. 963-968.

[25] H. Dai, Y. Wang, J. Fan, and B. Liu, "Mitigate ddos attacks in ndn by interest traceback," Proc. IEEE INFOCOM NOMEN Workshop, NJ, USA, 2013, pp. 22-29.

[26] N. Fotiou, P. Nikander, D. Trossen, and G. C. Polyzos, "Developing information networking further: From PSIRP to PURSUIT," Proc. 7th BROADNETS, Oct. 2010, pp. 22-27.

[27] P. Jokela, A. Zahemszky, C. E. Rothenberg, S. Arianfar, and P. Nikander, "LIPSIN: line speed publish/subscribe inter-networking," Proc. ACM SIGCOMM Conference on Data Communication, Barcelona, Spain, 2009, pp. 195-206.

[28] L. Carrea, A. Vernitski, and M. Reed, "Optimized hash for network path encoding with minimized false positives," Computer Networks, vol. 58, 2014, pp. 180-191.

[29] B. A. Alzahrani, M. J. Reed, and V. G. Vassilakis, "Resistance against brute-force attacks on stateless forwarding in information centric networking," Proc. ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS), Oakland, California, USA, May 2015.

[30] C. E. Rothenberg, P. Jokela, P. Nikander, M. Särelä, and J. Ylitalo, "Self-routing denial-of-service resistant capabilities using in-packet Bloom filters," Proc. European Conference of Computer Network Defence (EC2ND), 2009, pp. 46-51.

[31] B. A. Alzahrani, V. G. Vassilakis, and M. J. Reed, "Selecting Bloom-filter header lengths for secure information centric networking," Proc. 9th IEEE/IET CSNDSP, 2014, pp. 628-633.

[32] B. A. Alzahrani, V. G. Vassilakis, and M. J. Reed, "Mitigating brute-force attacks on Bloom-filter based forwarding," Proc. Conference on Future Internet Communications (CFIC), 2013, pp. 1-7.

 ISBN: 978-1-61208-411-4

Low Computational Design of Large Transmit Array MIMO Using Flexible Subarray Grouping

Tetsuki Taniguchi and Yoshio Karasawa
The University of Electro-Communications (UEC)
Tokyo, 182-8585 Japan
email: taniguch@ee.uec.ac.jp, karasawa@radio3.ee.uec.ac.jp

***Abstract*—This paper presents an improved version of low computational block diagonalization for Multiple Input Multiple Output (MIMO) multiuser telecommunication system equipped with a large array in transmitter side. While uniform subarray structure simply based on antenna index is adopted in previous work, this study investigates the effective flexible subarray grouping based on channel condition between antenna elements and receivers. In subarray grouping of large transmit array, all antenna elements are first sorted in the descent order of a certain metric, and then, they are grouped into subarrays according to a certain rule (in this paper, two metrics and two rules are considered). Through computer simulations, it is shown that this scheme is not effective for uniform subarray, but in nonuiform construction, it could achieve performance improvement under certain conditions.**

Keywords–Large array antenna; massive Multiple Input Multiple Output (MIMO); multiuser; subarray; zero-forcing.

I. INTRODUCTION

Multiple Input Multiple Output (MIMO) telecommunication system which utilizes array antenna in both of the transmitter and receiver sides has been established as a strategy which increases the data rate and/or reliability in the wireless communications [1]. To cope with the further increasing demand of the capacity, recently collecting attentions is so-called large or massive MIMO [2] which is equipped with large array one or both ends. It is actively studied also in industrial sector as given in [3]-[5]. In this system, the main topic is how to reduce the heavy computational load originated from a large number of antenna elements. For this aim, many approaches based on nonlinear processing [6] like suboptimal search method (for example, the application of tabu search in uplink for maximal likelihood detection [7], and in downlink for vector perturbation [8]) and belief propagation [9] have been proposed.

In the previous work [10], paying attentions to the fact that the linear processing approach which is popular in the conventional MIMO multiuser communication is not well investigated for the large array antenna model, we have proposed low computational version of block diagonalization for the MIMO system with large transmit array. This system first divide a large array into subarrays, and then, block diagonalization is applied to each of them: in [10], the subarrays are uniformly grouped simply based on antenna index not considering the state of the transmit antenna element to user connection. This study investigates the effect of more sophisticated grouping based on the channel condition between transmit antenna element to user, where two types of ordering metrics and two types of grouping rules are considered. Intuitively, subarray grouping based on the connection strength seems to bring performance improvement, but as shown later, this prediction does not consist in simple case like [10]. Then, when it is effective? How much is the improvement? The answers of those questions and features of flexible subarray grouping are shown in this paper.

The overall organization is as follows: Section II gives the system model of the multiuser MIMO considered in this study. Section III describes the low computational design method of the multiuser MIMO with large transit array utilizing the flexible subarray grouping. In Section IV, evaluation of the system performance is carried out through computer simulations based on the comparison with the conventional methods. Finally, in Section V, the conclusions of this study are described together with the future works.

II. SYSTEM MODEL

The model of MIMO multiuser system considered in this study is depicted in Figure 1, which consists of a transmitter Tx with N_t antennas and M receivers $\mathrm{Rx}_0, \cdots, \mathrm{Rx}_{M-1}$ with $N_{r,0}, \cdots, N_{r,M-1}$ antennas, respectively. Transmitter Tx transmits L_m data streams $\{s_{m,0}(t), \cdots, s_{m,L_m-1}(t)\}$ to receiver Rx_m using N_t-dimensional weight vectors $\boldsymbol{w}_{t,m,0}, \cdots, \boldsymbol{w}_{t,m,L_m-1}$. After passing propagation channel represented by $N_{r,m}$-by-N_t matrix H_m, a replica $\hat{s}_{m,\ell}(t)$ of data stream $s_{m,\ell}(t)$ is produced at Rx_m using $N_{r,m}$-dimensional weight vector $\boldsymbol{w}_{r,m,\ell}$. Here, the (n_r, m_t)-th element of H_m is the channel response between the n_t-th element of transmit array and the n_r-th element of receive array.

Though the picture of Figure 1 shows a typical MIMO multiuser downlink model, here, one assumption is added: the transmitter Tx has a large number of antenna elements (namely, N_t is a large number: in this study, $N_t = 128$) as shown in the simulation section (on the other hand, $N_{r,m}$ is assumed to be a small number: in this study, $N_{r,m} = 2$). Under this condition, generally, a heavy computational load is required for the design of the transmit weight, and the reduction of computation is an important topic in this kind of system. One solution is given in [10], and the next section suggests its improved version attempting better performance.

III. SYSTEM DESIGN

In the previous work in [10], we have presented a low computational design method of MIMO downlink system with a large transmit array, based on block diagonalization [6], which is a popular linear processing technique of multiuser MIMO. This section describes an improved version with flexible subarray grouping taking into account the channel condition.

First, a large transmit antenna is divided into S subarrays $\mathrm{Tx}_0, \cdots, \mathrm{Tx}_{S-1}$. Subarray Tx_s consists of antennas with index set $\mathcal{N}_t^{(s)} = \{n_{t,s,0}, \cdots, n_{t,s,N_{t,s}-1}\}$, where any

 ISBN: 978-1-61208-411-4

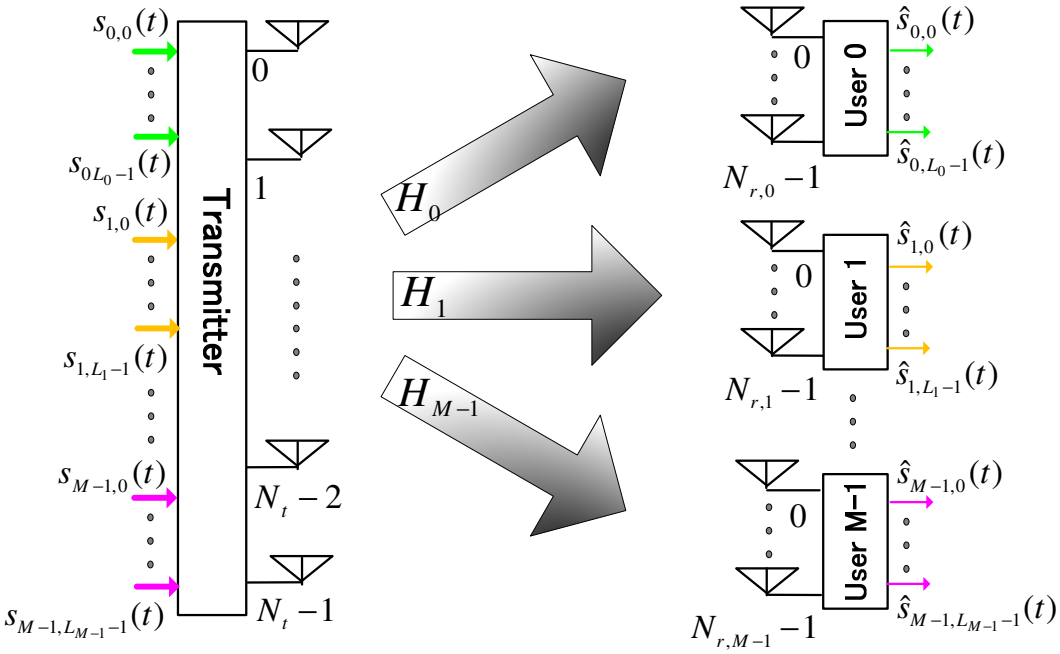


Fig. 1 Model of MIMO multiuser system (downlink).

antenna element belongs to only one of S subarrays (therefore, $N_t = \sum N_{s,k}$ consists). Here, the problem is how to construct those subarrays. Though uniform subarrays are constructed simply in the order antenna index regardless of the channel state in [10], namely, $\mathcal{N}_t^{(s)} = sN_{t,0} + \{0, \cdots, N_{t,0} - 1\}$ (remark that $N_{t,s} = N_{t,0}$ for any s), in this study, we consider flexible subarray grouping. The whole procedure of flexible subarray grouping is depicted in Figure 2.

In the first step, one of the following two metrics based on the channel condition between the n_t-th transmit antenna element and the m-th receiver end is calculated:
(M-1) *Norm Metric*: This metric is defined by $f_{n_t} = \|\boldsymbol{h}_{m,n_t}\|$, where $\boldsymbol{h}_{m,n_t}$ is a Single Input Multiple Output (SIMO) channel defined by $\boldsymbol{h}_{m,n_t} = [H_{m,0,n_t}, \cdots, H_{m,N_{r,m}-1,n_t}]^T$, and H_{m,n_r,n_t} is a scalar which shows response between the n_t-th antenna element of Tx and the n_r-th antenna element of Rx_m.
(M-2) *Absolute Sum Metric*: This metric is defined by equation $f_{n_t} = \sum_{n_r} |H_{m,n_r,n_t}|$, which does not require any multiplication operation. It is more preferable choice compared with norm metric from the viewpoint of the computational cost (remark that since the calculation of metric is carried out user by user, advantage of this simplicity becomes large).
In the following, the transmit antenna elements are renumbered in descent order, namely, indices $\{n_{t,0}\, n_{t,1} \cdots, n_{t,N_t-1}\}$ such that $f_{n_{t,0}} \geq f_{n_{t,1}} \geq \cdots \geq f_{n_{t,N_t-1}}$ are replaced by new indices $\{0, 1, \cdots, N_t - 1\}$ (after this operation, $f_0 \geq f_1 \geq \cdots \geq f_{N_t-1}$ consists). The conventional approach in [10] can be regarded as a scheme adopting random number as metric f_{n_t} because of the i.i.d. (independent and identically distributed) statistics of the fading.

In the second step, based on the above ordering, subarray grouping is carried out according to one of the following two rules:

(R-1) *Even Grouping*: Antenna $0, 1, 2, \cdots$ are grouped into $\mathcal{N}_t^{(0)}, \mathcal{N}_t^{(1)}, \mathcal{N}_t^{(2)}, \cdots$ in the order of the antenna index after the renumbering. If it reached to $s = S - 1$ (namely, $\mathcal{N}_t^{(S-1)} = \{S - 1\}$), then the subarray index return to 0, like $\mathcal{N}_t^{(0)} = \mathcal{N}_t^{(0)} \cup \{S\}$, $\mathcal{N}_t^{(1)} = \mathcal{N}_t^{(1)} \cup \{S + 1\}$, $\cdots$. If the number of elements in $\mathcal{N}_t^{(s)}$ has reached to $N_{t,s}$, this subarray is excluded from the above circulation. This operation attempts to keep the average metric of all the subarrays equal as possible.
(R-2) *Uneven Grouping*: Transmit antennas are grouped like $\mathcal{N}_t^{(s)} = \sum_{k=0}^{s-1} N_{t,k} + \{0, \cdots, N_{t,s}\}$, namely, $N_{t,0}$ antennas with the strongest metric are grouped into the first subarray, and those with the next strongest $N_{t,1}$ are distributed to the second subarray. This operation is repeated all the N_t elements are grouped into one of subarrays.

As a consequence of the above subarray grouping, subchannel between subarray Tx_s and receiver Rx_m is derived as $H_m^{(s)} = [\boldsymbol{h}_{m,n_{t,s,0}}, \cdots, \boldsymbol{h}_{m,n_{t,s,N_{t,s}-1}}]$.

After subarray construction, the conventional block diagonalization is applied to each subarray. Since the optimal receiver weight is different subarray by subarray, two methods – Method 1 (Tx and Rx simultaneous design) and Method 2 (Rx first Tx second design) – are considered to determine it uniquely (for the detail, see [10]).
(i) *Method 1*: Subarray weight vector $\boldsymbol{w}_{t,m}^{(s)} = V_m^{(s)} \boldsymbol{v}_m^{(s)}$ is calculated by block diagonalization of $H_m^{(s)}$, where the columns of $V_m^{(s)}$ span the kernel of $H_{-m}^{(s)} = [H_0^{(s)T}, \cdots, H_{m-1}^{(s)T}, H_{m+1}^{(s)T}, \cdots, H_{M-1}^{(s)T}]$, and $\boldsymbol{v}_m^{(s)}$ is the right singular value vector of $H_m^{(s)} V_m^{(s)}$ corresponding to the largest singular value. Then, virtual channel $H_{v,m} = [H_m^{(0)} \boldsymbol{w}_{t,m}^{(0)}, \cdots, H_m^{(S-1)} \boldsymbol{w}_{t,m}^{(S-1)}]$ is cal-

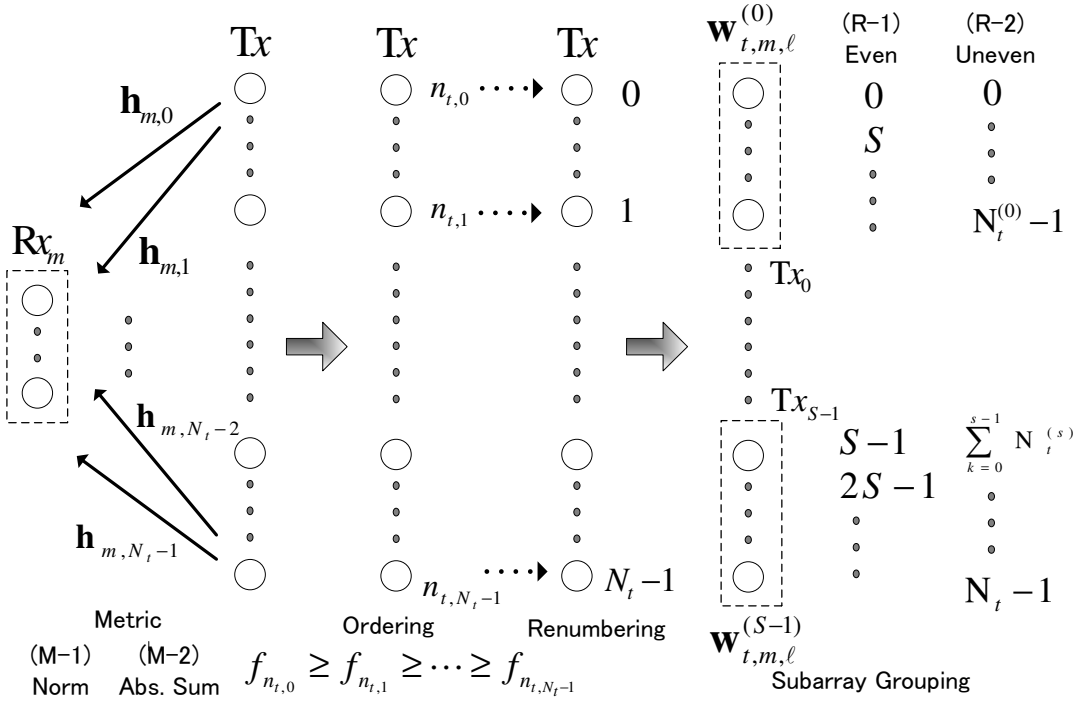


Fig. 2 Procedure of flexible subarray grouping.

culated, and transmitter and receiver weights $\boldsymbol{w}_{t,m,\ell} = [c_{m,\ell}^{(0)}\boldsymbol{w}_{t,m}^{(0)T}, \cdots, c_{m,\ell}^{(S-1)}\boldsymbol{w}_{t,m}^{(S-1)T}]^T$ and $\boldsymbol{w}_{r,m,\ell}$ are derived, where $\boldsymbol{w}_{r,m,\ell}$ and $\boldsymbol{c}_{m,\ell} = [c_{m,\ell}^{(0)}, \cdots, c_{m,\ell}^{(S-1)}]^T$ are the left and right singular value vectors of $H_{v,m}$ corresponding to the ℓ-th largest singular value (it is defined as $\ell = 0$ for the largest one).

(ii) *Method 2*: Receiver weight $\boldsymbol{w}_{r,m,\ell}$ is derived as the left singular value vector of H_m, and then, channel $\tilde{H}_m = [H_m^T\boldsymbol{w}_{r,m,0}^*, \cdots, H_m^T\boldsymbol{w}_{r,m,L_m-1}^*]^T$. After the calculation of $\boldsymbol{w}_{t,m}^{(s)}$ and $H_{v,m}$ is carried out in a similar manner as Method 1 but using $\tilde{H}_m$ instead of H_m, vector $\boldsymbol{c}_{m,\ell} = \boldsymbol{h}_{m,\ell}$ is derived from $H_{v,m} = [\boldsymbol{h}_{m,0}^T, \cdots, \boldsymbol{h}_{m,L_m-1}^T]^T$.

The additional computational cost introduced by the proposed method is generally small. Once the metric is given, that for the distribution to groups is almost negligible. For the calculation of metrics, $N_t \sum N_{r,m}$ multiplications should be performed under the use of norm metric, but if the absolute sum metric with no multiplication suggested in this paper (the performance degradation from the norm metric is very small) is used, the complexity is only slightly increased, and not so much different compared with that of BD.

IV. Computer Simulations

In this section, computer simulations are carried out to verify the effectiveness and features of the proposed method described in Section III. The default simulation conditions are summarized in Table I.

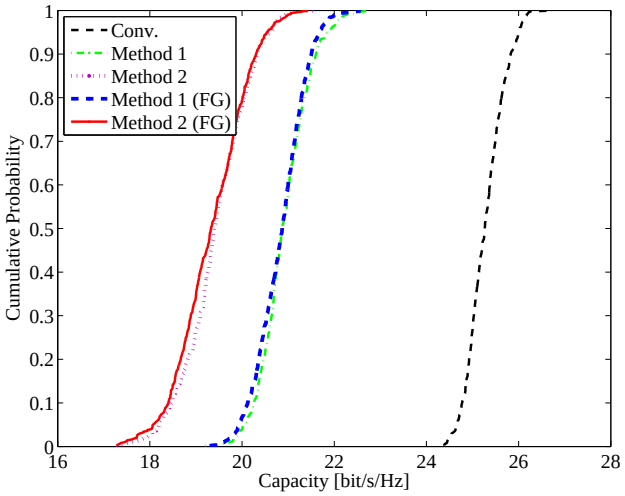
The performance is evaluated using sum capacity represented by $C_m = \sum_\ell \log_2(1 + \mathrm{SINR}_{m,\ell})$ for the m-th user, where $\mathrm{SINR}_{m,\ell}$ is the Signal to Interference plus Noise Ratio (SINR) of the output signal $\hat{s}_{m,\ell}(t)$ of the ℓ-th stream of the m-th user (it is calculated by using 200 samples of fading channels). On the other hand, Signal to Noise Ratio (SNR) is defined by $\mathrm{SNR}_m = P_{s.m}/P_{n,m}$, where $P_{s,m} = 1$ and $P_{n,m}$ are the total energy of the transmitted signal and the receiver noise.

The number of users is $M = 4$ and each of them is equipped with $N_{r,m} = 2$ antennas, and two data streams are send to Rx_m (hence $L_m = 2$) using $N_t = 128$ antennas from transmitter Tx based on the energy allocation by the

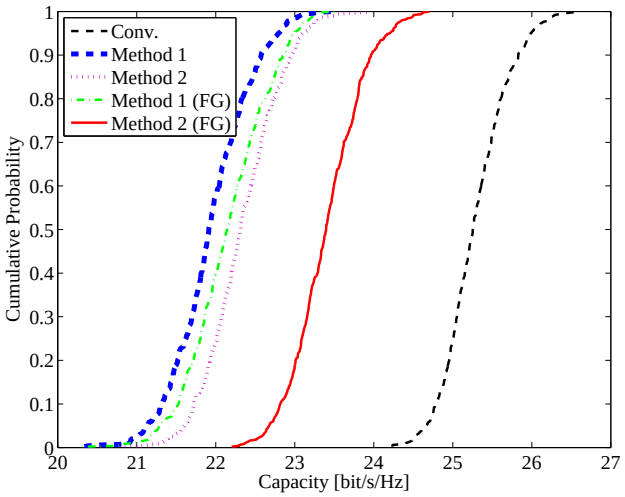
TABLE I. Simulation Conditions

Number of Receivers (Number of Users)	$M = 4$
Number of Transmit Antennas	$N_t = 128$
Number of Receive Antennas	$N_{r,m} = 2$
Number of Streams per User	$L_m = 2$
Subarrays (uniform)	$S = 16$ $N_{t,s} = 8$ $\mathcal{N}_t^{(s)} = 8s + \{0, \cdots, 7\}$ $s = 0, \cdots, S-1$
Subarrays (nonuniform)	$S = 12$ $N_{t,s=0\sim3} = 16$ $\mathcal{N}_t^{(s=0\sim3)} = 16s$ $+\{0, \cdots, 15\}$ $N_{t,s=4\sim S-1} = 8$ $\mathcal{N}_t^{(s=4\sim S-1)} = 8(s-4)$ $+\{0, \cdots, 7\} + 64$
Grouping Metric	Absolute Sum
Grouping Rule	Uneven Grouping
Energy Constraint	$P_m = 1$
SNR	$\mathrm{SNR}_m = 20\mathrm{dB}$
Channel Statistics	i.i.d. Rayleigh Fading with unit variance

 ISBN: 978-1-61208-411-4



(a) Uniform subarray grouping.



(b) Nonuniform subarray grouping.

Fig. 3 Distribution function of capacity.

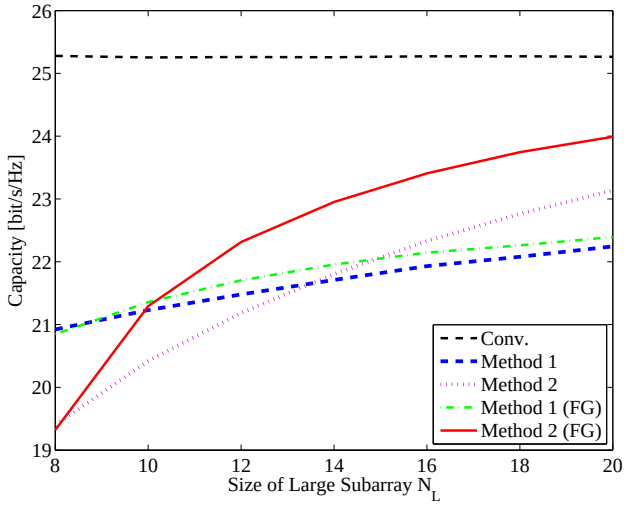


Fig. 4 Large subarray size versus capacity.

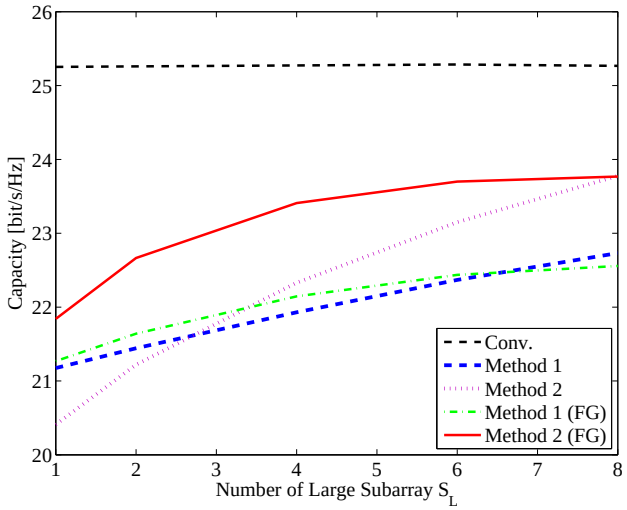


Fig. 5 Number of large subarray versus capacity.

water filling. Under this condition, default size and number of subarray are $N_{t,s} = 8$ and $S = 16$ for uniform case, and $N_{t,s=0\sim3} = 16$, $N_{t,s=4\sim S-1} = 8$ and $S = 12$ for nonuniform case.

Figure 3 depicts the distribution functions of capacity (since the curves of all users are almost same because of the symmetry of fading statistics in channels, it is drawn only for one user) for the case of the absolute sum metric and uneven grouping. Five curves respectively shows the results of the conventional block diagonalization (Conv.), Method 1 and Method 2 in [10], and those with flexible subarray grouping in Section III (Method 1 (FG) and Method 2 (FG)). In subplot (a) for uniform subarray grouping, the curves of Method 1 (FG) and Method 2 (FG) are overlapped with those of Method 1 and Method 2, which means flexible grouping is *not effective*. On the contrary, in nonuniform subarray grouping in subplot (b), the performance of Method 1 (FG) and Method 2 (FG) is better than that of Method 1 and Method 2, which means flexible grouping is *effective* (in this subplot, the position of the curves of Method 2 and its FG version is moved to the right of those of Method 1 and its FG version, since the increased degrees of freedom in a large subarray brings advantage to Method 2 [10]). From those results, it is considered that the proposed flexible grouping approach is a better choice when nonuniform array is adopted and antennas with the strong connection to the user is grouped into large subarrays. Another important point is that the performance improvement between Method 2 and Method 2 (FG) is larger than that of Method 1 and Method 1 (FG).

Here, though the graph is not shown, the results of the norm and absolute sum metrics have only slight difference, hence we adopt absolute sum metric with much smaller computation in the rest of the simulation. We have also verified that the even grouping does not bring the performance improvement regardless of subarray construction. The reason is considered as follows: because of the randomness of the fading channel, even without antenna ordering, the average metric becomes almost same among subarrays.

To investigate the influence of the subarray size in the proposed approach, Figure 4 plots the large subarray size N_L

 ISBN: 978-1-61208-411-4

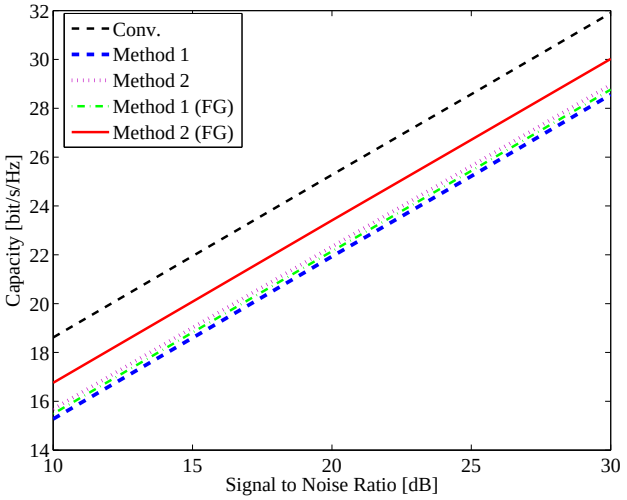


Fig. 6 SNR versus capacity.

versus capacity relation for nonuniform subarray construction. Among S subarrays, four subarrays $\mathrm{Tx}_{s=0\sim3}$ have large size $N_L = N_{t,s=0\sim3} \in \{8,\ 10,\ 12,\ 14,\ 16,\ 18,\ 20\}$, and other $S-4$ subarrays have short size $N_S = N_{t,s=4\sim S-1} = 8$. According to the uneven grouping rules, the antennas with strong connection to the target receiver are gathered into large subarrays. From this figure, it can be verified that the performance improvement by the proposed method becomes larger as the large subarray size in nonuniform scheme is increased. In this figure, Method 2 (FG) has better performance than the conventional methods in [10] except the uniform case of $N_S = N_L = 8$. It is also shown that the advantage of Method 2 (FG) against Method 1 (FG) is large under the large subarray size.

Figure 5 depicts the number of large subarray S_L versus capacity curves. Among S subarrays, $\mathrm{Tx}_{s=0\sim S_L-1}$ have large size $N_L = N_{t,s=0\sim S_L-1} = 16$, and other $S_S = S - S_L$ subarray have short size $N_S = N_{t,s=S_L\sim S-1} = 8$. What can be observed from this picture is that the performance improvement by the proposed method compared with the conventional approach [10] becomes small as the number of large subarray increases. This result shows the proposed method is advantageous when the uniformity of the subarray construction is small.

The relation between SNR and per-user capacity is drawn in Figure 6. From this figure, it can be seen that the relation among five methods does not depend on SNR.

V. Conclusions and Future Works

This paper has investigated on flexible subarray grouping in low computational block diagonalization for MIMO multiuser system adopting a large array in the transmitter side. All transmit antenna elements are first sorted in the descent order of one of two metrics (norm and absolute sum), and then, they are grouped into one of subarrays according to one of two rules (even and uneven grouping). Computer simulations have demonstrated that the proposed approach is useless for uniform subarray construction, but as the nonuiformity of subarray increases, one of two rules, uneven strategy invokes its effectiveness under certain conditions, and the performance improvement compared with the conventional approach in [10] becomes larger, where also shown is the simple absolute sum metric is sufficient.

A future work is the investigation on the subarray processing based on Minimum Mean Square Error (MMSE) criterion. Utilization of the subarray strategy for the signal detection in the large array is another attractive topic of study.

Acknowledgment

This research was supported in part by the Grant-in-Aid for Scientific Research (No. 25249056) from Japan Society for the Promotion of Science (JSPS).

References

[1] Y.-S. Cho, J. Kim, W.-Y. Yang, and C.-G. Kang, MIMO-OFDM Wireless Communications with MATLAB, Wiley, 2010.

[2] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," IEEE Commun. Mag., vol. 52, no. 2, Feb. 2014, pp. 186-195.

[3] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: how many antennas do we need?" IEEE J. Selec. Areas Commun., vol. 31, no. 2, Feb. 2013, pp. 160-171.

[4] C. Kim, T. Kim, and J.-Y. Seol, "Multi-beam transmission diversity with hybrid beamforming for MIMO-OFDM systems," 2013 IEEE Globecom Workshops, Dec. 2013, pp. 61-65.

[5] S, Akoum and J. Acharya, "Full-dimensional MIMO for future cellular networks," 2014 IEEE Radio and Wireless Symposium (RWS 2014), Jan. 2014, pp. 1-3.

[6] Q. H. Spencer, C. B. Peel, A. L. Swindlehurst, and M. Haardt, "An introduction to the multi-user MIMO downlink," IEEE Commun. Mag., vol. 42, no. 10, Oct. 2004, pp. 60-67.

[7] Y. Wu and J. McAllister, "FPGA-based Tabu search for detection in large-scale MIMO systems," 2014 IEEE Workshop Signal Process. Syst. (SiPS 2014), Oct. 2014, pp. 1-6.

[8] W. Ding, L. V. Tiejun, A. Hu, and Sixi Su, "A low-complexity vector precoding scheme for large multiuser MIMO systems," 2013 16th Int. Symp. Wireless Personal Multimedia Commun. (WPMC 2013), June 2013, pp. 1-5.

[9] S. Wu, L. Kuang, Z. Ni, J. Lu, D. D. Huang, and Q. Guo, "Low-complexity iterative detection for large-scale multiuser MIMO-OFDM systems using approximate message passing," IEEE J. Selec. Topics Signal Process, vol. 8, no. 5, Oct. 2014, pp. 902-915.

[10] T. Taniguchi and Y. Karasawa, "A low computational design of multiuser MIMO system using large scale array antenna," 2014 5th Int. Conf. Network Future (NOF 2014), Dec. 2014, pp. 1-6.

 ISBN: 978-1-61208-411-4

Sparse Construction of Joint Viterbi Detector Decoder (JVDD) Codes

Ashish James, Kheong Sann Chan, and Sari Shafidah Binte Shafee
Data Storage Institute (DSI), A*STAR (Agency for Science Technology and Research)
Singapore 117608
Email: {jamesa, CHAN_Kheong_Sann, Sari_S}@dsi.a-star.edu.sg

***Abstract*—The Joint Viterbi detector decoder (JVDD) has been proposed as an alternative to the iterative detector, performing both detection and decoding in two stages on a trellis. The first stage estimates and retains a set of survivors, while the second stage performs a parity check on these to compute the minimum metric legal codeword (MMLC). With this structure, near optimal maximum-likelihood decoding (MLD) performance can be achieved but at the cost of complexity especially at long codeword lengths (CWL). JVDD codes have been introduced with the explicit target of reducing this complexity. Further, lower rate codes with more parity checks leads to reduced number of survivors in the JVDD trellis resulting in lower complexity. However, it has been observed that JVDD code performance degrades at low-rates while operating in the low SNR region through an error-floor. This aspect is analyzed in this paper and this performance can be attributed to the JVDD code structure where a constant number of ones are placed to the left of the main diagonal. This results in a dense region of ones at the top left corner where the entire rows are filled with ones. In this paper, a modification to the JVDD code construction is proposed by adjusting the number of ones placed in those rows where the row weight of JVDD codes is less than the available row width. It has been found that such sparse construction results in reducing JVDD complexity, as well as eliminates the error-floor problem. This has been verified through extensive simulation studies.**

Keywords–JVDD; Iterative detector; ISI Channel; Sparse construction of JVDD Codes.

I. INTRODUCTION

Iterative detectors and decoders have been a subject of intense research due to their outstanding error correcting capabilities with performance very close to the Shannon limit. Although iterative decoding can achieve channel capacity as the block size goes to infinity, there is still a gap with the optimal maximum-likelihood (ML) decoder for any code-structure [1]. ML decoders have been analyzed over different communication channels - additive white Gaussian noise (AWGN) [2] [3], binary symmetric channel (BSC) [3], binary erasure channel (BEC) [4] among others. These have also been employed for two broad classes of codes namely block codes and convolutional codes. With convolutional codes, efficient trellis based algorithm known as Viterbi algorithm (VA) can be employed to perform ML decoding and return the most probable transmitted codeword [5]. However, optimal ML decoding of linear block codes has been proven to be an NP-hard problem [6], whose complexity grows exponentially as the code length increases. There have been many research efforts in this direction to develop optimal or suboptimal decoding algorithms with moderate complexity [7]–[11].

In [11], they introduce the joint Viterbi detector decoder (JVDD) as an alternative optimal ML detection and decoding scheme that attempts to return the minimum metric legal codeword (MMLC) . It operates on a trellis and has a two-stage decoding structure - metric thresholding and parity checking. The first stage executes the normal VA by computing metrics for every possible path to a node. However, the JVDD retains the minimum metric survivor along with a certain number of competing paths in the trellis constrained by a threshold parameter. Thus only survivors with metrics within the threshold of the minimum metric for a particular node are retained. This would typically mean setting a larger threshold to minimize the probability of discarding the MMLC. However, this leads to a larger number of survivors resulting in an increased complexity. The second stage, aims at reducing the complexity by performing parity checking on each incoming survivor path and discarding those which fail the syndrome check, i.e., $c\mathbf{H}^T = 0$, where c is the codeword and $\mathbf{H}$ is the parity check matrix. This parity checking section provides a system tradeoff design between complexity and code-rate. Thus JVDD performance complexity can be reduced by increasing the number of parity checks per bit, i.e., by operating at low code-rates. However, it is found that JVDD codes exhibits an error-floor at low code-rates while operating in the low SNR region. This is the focus of the present paper. The error floor is characterized by a more gradual decrease in error rate as code-rate decreases and can be attributed to the JVDD code structure.

JVDD codes are currently designed with a parity check matrix that has constant number of ones in a row (row weight) placed according to a Gaussian distribution to the left of the diagonal. However, this construction results in a dense region of ones towards the top left corner where the entire row gets filled with ones where the length of the row (row width) is lower than the row weight as observed in Figure 1 (In this figure, the 1's in the parity check matrix are represented as black dots while white spaces represent the 0's). This paper analyses the impact of this dense region through a modification of this construction by adjusting the row weight in this region of the parity check matrix. The results indicate that the error-floor can be mitigated by eliminating this dense region of ones.

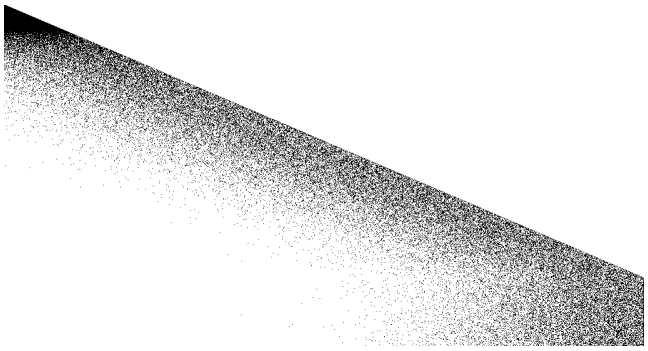


Figure 1. Pictorial depiction of the parity check matrix $\mathbf{H}$ of JVDD codes for codeword length of 1024 and 0.5 rate.

The rest of the paper is organized as follows. Section II describes the JVDD algorithm along with sparse construction of JVDD codes. Simulation results are presented in Section III, followed by concluding remarks in Section IV.

II. SPARSE CONSTRUCTION OF JVDD CODES

In this paper, binary linear block coding for BPSK signalling over an inter-symbol interference (ISI) channel is considered. Then, the received signal y_k at time k is given as

$$y_k = \sum_{l=0}^{L} f_l x_{k-l} + w_k \quad (1)$$

 ISBN: 978-1-61208-411-4

where f_l is the channel impulse response of order L, x_k is the transmitted encoded bit sequence and w_k corresponds to AWGN with zero mean and variance N_0.

A. JVDD Codes

The receiver is implemented based on the JVDD algorithm, which operates on a trellis where the paths through the trellis correspond to the codewords c that satisfy the parity check condition: $c\mathbf{H}^T = 0$. However, the received sequence y_k may not correspond to a codeword due to the channel and the JVDD finds the path through the trellis which is the closest to the received sequence y_k that is also a legal codeword. This corresponds to the maximum likelihood criteria represented as

$$\max_{c \in C} \sum_{k=0}^{N-1} \ln \Pr(y_k|c_k) = \min_{c \in C} \sum_{k=0}^{N-1} \gamma(y_k|c_k) \tag{2}$$

where $\gamma(y_k|c_k)$ is the branch metric. The branch metrics corresponds to the weights of the trellis transitions and the calculation for each possible transition from state i to l at time step k is given as

$$\gamma_{k,k+1}^{j,l} = c_k y_k = \begin{cases} y_k & \text{if } i \neq l \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

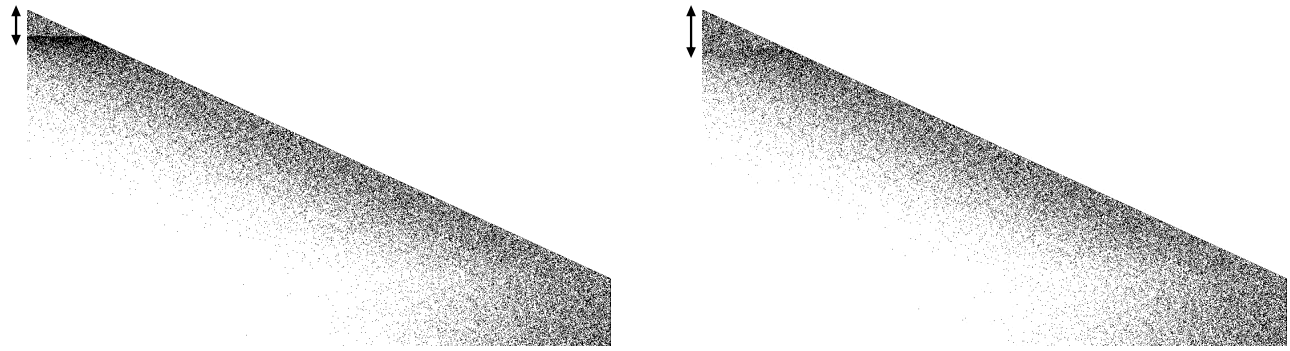
For each state s_{k+1} at time $k+1$, the state metric $\alpha_{l,k+1} = \min(\alpha_{l,k}, \alpha_{i,k} + \gamma_{k,k+1}^{j,l})$ are calculated and correspond to the survivor paths. In the usual Viterbi algorithm, the metrics for each incoming survivor to a node are computed and the larger-metric survivors are discarded. However, this might result in discarding the MMLC and deteriorate the result of subsequent detections. To resolve this problem, the JVDD adjusts the number of surviving paths through metric thresholding, which computes the path metrics but discards survivors with metrics larger than the *threshold*. Let τ denote the threshold, which determines the number of competing paths that are retained for each state to enhance the information update process. Accordingly, the survivors with state metric $\alpha_{l,k+1} < \tau + \alpha_{min}$ (where α_{min} is the minimum metric) are retained. This would typically mean setting a larger threshold for minimizing the probability of discarding the MMLC. However, this leads to a larger number of survivors resulting in an increased complexity.

The complexity of the JVDD by retaining survivors in the metric thresholding section can be reduced through the parity checking stage that follows. Parity checking occurs on specific nodes corresponding to the last one of a given row of the parity check matrix. At these particular nodes, all bits required to perform the check are detected and the syndrome can be computed. In this stage, the JVDD performs parity checks on each incoming survivor path to the parity check node and discards paths which fail the syndrome check, i.e., $ch_i^T = 0$ where c is the detected codeword and h_i^T is the transpose of the ith row of the parity check matrix $\mathbf{H}$. The parity check matrix $\mathbf{H}$, is an $M \times N$ matrix where M is the number of parity checks and N is the number of coded bits, with each row corresponding to one of the M parity checks. In this context, JVDD codes were designed to evenly space the parity checking functionality throughout the trellis resulting in fewer number of survivors especially as the codeword length (CWL) increases. This is achieved by designing the parity check matrix with a constant number of ones in a row (row weight) placed according to a Gaussian distribution to the left of the main diagonal. However, this construction results in a dense region of ones towards the top left corner where the entire row gets filled with ones as the length of the row (row width) is lower than the row weight as observed in Figure 1. This paper analyses the impact of this dense region through a modification of this construction as described in the following section.

B. Sparse JVDD Codes

Since the JVDD codes are generated using a parity check matrix with a fixed row weight, traditional JVDD code construction results in a dense region of ones when the row width is less than the row weight. This aspect is addressed in this paper through the sparse JVDD code construction by adjusting the row weight in this region of the parity check matrix $\mathbf{H}$. In sparse JVDD code construction, instead of having a constant row weight, the number of 1's placed in $\mathbf{H}$ is varied according to the width of the row. This paper considers two levels of depth for sparseness - region where the row width is less than or equal to the row weight and region where the row width is equal to 1.5 times the row weight. The number of 1's in these rows is placed randomly based on Gaussian distribution to the left of main diagonal and is a factor of the row width. This results in a less dense $\mathbf{H}$ matrix where the rows till the depth of sparseness is filled with only a specified percentage of 1's based on the row width and is referred to hereby as the level of sparseness. Figure 2 is a pictorial representation of a parity check matrix generated through such a construction with 50% sparseness, i.e., filling 50% of the row width with 1's.

The effect of such a sparse construction on the column weight of parity check matrix is shown in Figure 3. It is seen



(a) Depth - Row width $\leq$ Row weight

(b) Depth - Row width $\leq 1.5\times$Row weight

Figure 2. Pictorial depiction of the parity check matrix $\mathbf{H}$ of Sparse JVDD codes for codeword length of 1024 and 0.5 rate at different depth of sparseness. 1's are represented as black dots while white spaces represent 0 in these figures.

 ISBN: 978-1-61208-411-4

that sparse construction results in lowering the column weight for the depth of sparseness.

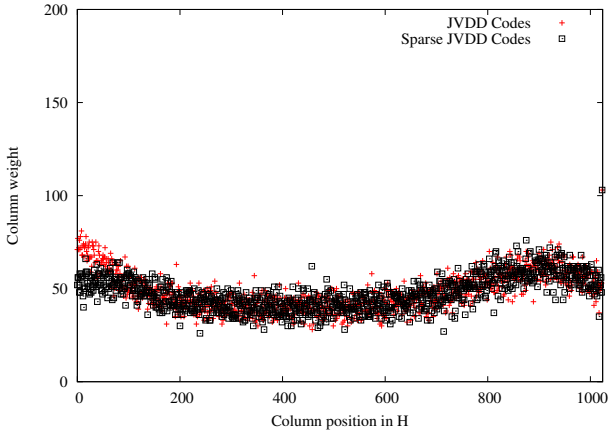


Figure 3. Variation of column weight of parity check matrix **H** of Sparse and traditional JVDD codes for codeword length of 1024.

The class of JVDD codes used in this paper is the variable-gradient Gaussian distribution linear diagonal (VGGDLD) codes introduced in [12], where the gradient of the diagonal is varied through two independent shift parameters dx and dy. As observed in [12], increasing the parameter dx leads to increased number of survivors in the trellis as the parity checking gets shifted in the trellis, thus the optimum value is set to $dx = 0$. However, the parameter dy has to be designed optimally, where too small a value leads to under-protected bits at the end of the codeword, and too large a value reduces the number of parity checks which increases the complexity of the trellis. The impact on performance and complexity of JVDD by implementing sparse construction on the VGGDLD codes is studied in this paper.

III. Simulation Results

In this section, performance results of sparse JVDD codes are analyzed. For simulations, the codeword length (CWL) is fixed at 1024, i.e., the number of columns N in the parity check matrix **H** is 1024. The number of rows M in the parity check matrix **H** is varied from 512 to 102 which correspond to code-rates 0.5 to 0.9. The best VGGDLD codes with typical values of dx=0% of N and dy=20% of M are employed for the simulations. Sparse construction is implemented on such codes and the level of sparseness is varied to determine the optimal sparseness for such codes. JVDD performance is also compared with iterative detector employing random codes at the same code-rates which are used as a benchmark. The simulation parameters are specified in Table I.

TABLE I. Simulation Parameters

Codeword Length (CWL)	1024
Code-rate (R)	[0.90, 0.85, 0.80, 0.70, 0.60, 0.50]
SNR (dB)	[6, 7]
[dx (**% of** N), dy (**% of** M)]	[0, 20]
Row weight	100
Iterative detector	Random code (column weight - 4)
JVDD Max. No. Survivors	30000
Channel (f)	$\frac{1}{\sqrt{6}}$[1 2 1]
Level of sparseness	[1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 99]

A. Complexity of JVDD with sparse codes

Initially, the complexity of the JVDD algorithm employing sparse codes is analyzed, where the complexity is measured as the average number of survivors in the JVDD trellis. This set of simulations are performed for a sparseness depth where the row weight is less than or equal to the row width. Figure 4 depicts the variation of complexity with threshold for a code-rate of 0.5 at an SNR of 6 dB. Typically JVDD performance enhances with threshold but at the cost of increasing complexity [11]. The level of sparseness is also varied and it is observed that the complexity increases when the JVDD codes are too sparse (1%) or highly dense (99%). This basically means an increase in the number of survivors when JVDD is processing that region. Thereby, it is desirable to produce codes that minimize JVDD complexity, as analyzed in the following section.

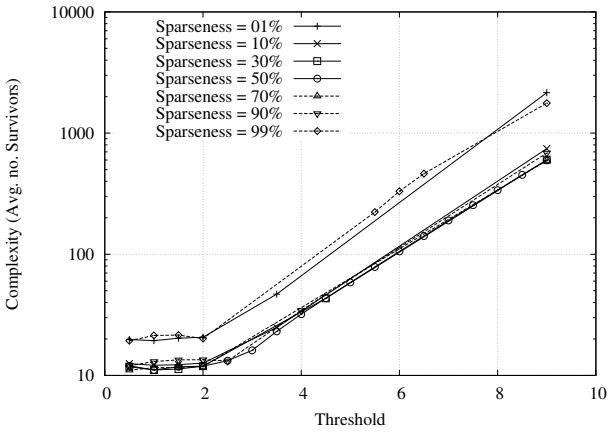


Figure 4. Variation of JVDD complexity with threshold and level of sparseness at SNR of 6 dB and 0.5 code-rate

In order to realize the optimal level of sparseness, the complexity of JVDD is analyzed against sparseness and frame error rate (FER) as depicted in Figure 5 and 6, respectively. Similar to the previous observation, too sparse and highly dense codes increases the complexity of the JVDD. However, the complexity is found to minimize and saturate when sparseness is in the range 30 - 70% as observed in Figure 5. Further, in Figure 6 the optimal level of sparseness to achieve the lowest FER is found to be 50%. Thereby from Figures 4, 5 and 6, typical level of sparseness considered for further analysis is 50%.

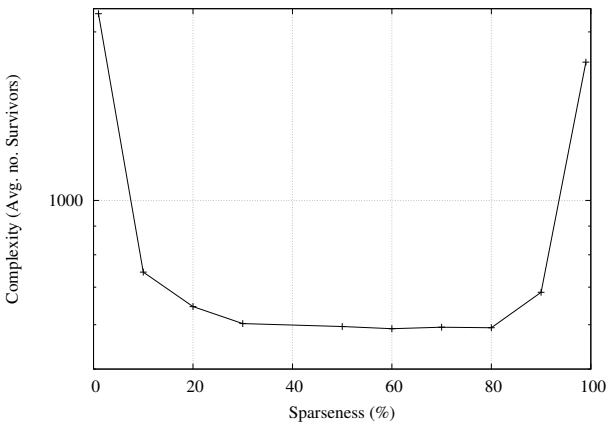


Figure 5. Variation of JVDD complexity with sparseness at SNR of 6 dB and 0.5 code-rate

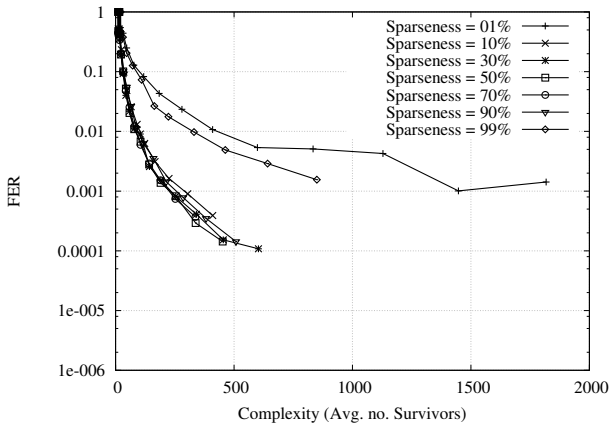


Figure 6. FER comparison with JVDD complexity at different level of sparseness at SNR of 6 dB and 0.5 code-rate

Figure 7 depicts the JVDD complexity at different depths of sparseness and varying code-rates at an SNR of 6 dB. The depth of sparseness refers to the number of rows where sparseness is introduced into the JVDD codes. Two levels are compared where sparseness is introduced in the region where row weight is less than or equal to - the row width and 1.5 times the row width. It is observed that the complexity remains almost the same irrespective of the depth of sparseness at different code-rates. Further it is shown that JVDD complexity increases with code-rate. This can be attributed to the fact that at lower code-rates, there are more parity checks per information bit which means that parity checking will occur more frequently in the JVDD trellis thereby keeping the number of survivors more manageable.

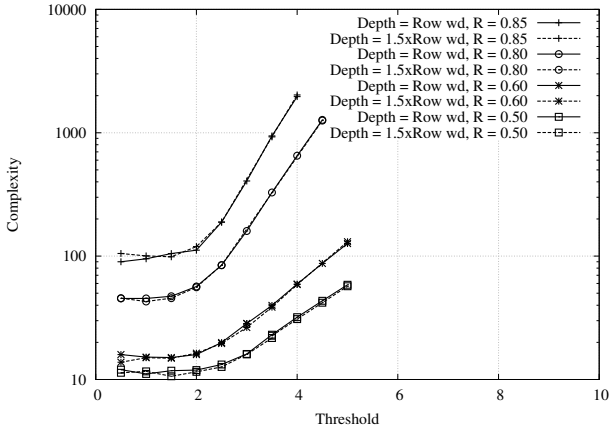
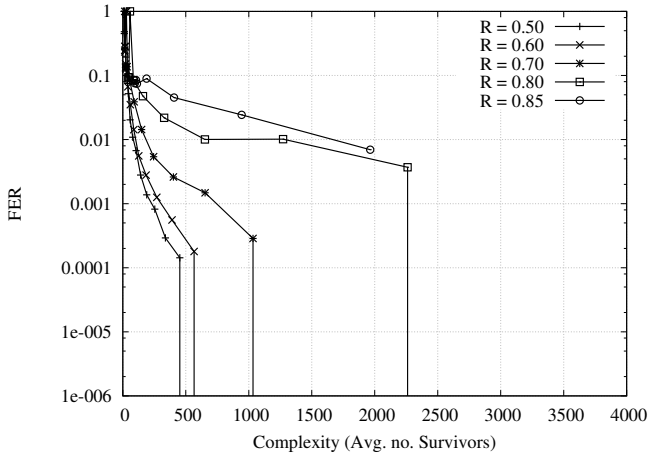
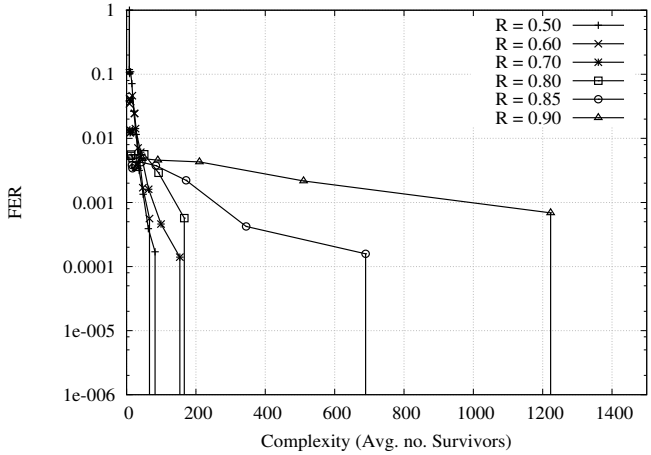


Figure 7. Variation of JVDD complexity with threshold at different depths of sparseness and code-rates at SNR 6 dB

Figure 8 shows the variation of FER with complexity of JVDD at different rates. The operating SNRs considered for comparison is 6 and 7 dB. It is observed that the complexity increases with code-rate and decreases with increasing SNR. Lowering the code-rate is expected to reduce errors in any channel at the cost of increased coding overhead through the increased number of parity check bits. For the JVDD, this also results in reducing the number of survivors as parity checking kills survivors that violate the syndrome. Further, increasing the SNR will result in reduced distortions from the channel and will lower the chance of deviating from the correct path in the JVDD trellis. This results in fewer survivors and reduced complexity. These trends indicate that operating at low code-rates would benefit JVDD due to reduced complexity but the original JVDD code performance is found to deteriorate in these conditions due to the dense region of ones in the upper left corner of the **H** matrix as observed in the following section.



(a) SNR = 6 dB



(b) SNR = 7 dB

Figure 8. FER comparison with JVDD complexity at different rates and SNRs.

B. FER Performance of Sparse JVDD Codes for Different Rates

Sparse JVDD code performance is compared with traditional JVDD codes and iterative detector at various code rates in Figure 9. It is known that JVDD code performance improves with increasing complexity and the performance comparison needs to be performed by normalizing the complexity. In this work, the complexity is normalized to 1000 and 2000 survivors, respectively. It is observed that the original JVDD code performance (with the dense region of ones) deteriorates at low code-rates through the appearance of an error floor which is characterized by a gradual decrease in error rate as code-rate decreases. This performance degradation has been mitigated through the sparse construction developed in this paper. This confirms that the performance loss of traditional JVDD codes can be attributed to the code construction which results in a dense region of ones towards the top left corner

 ISBN: 978-1-61208-411-4

where the entire row gets filled with ones as the length of the row (row width) is much lower than the row weight as observed in Figure 1.

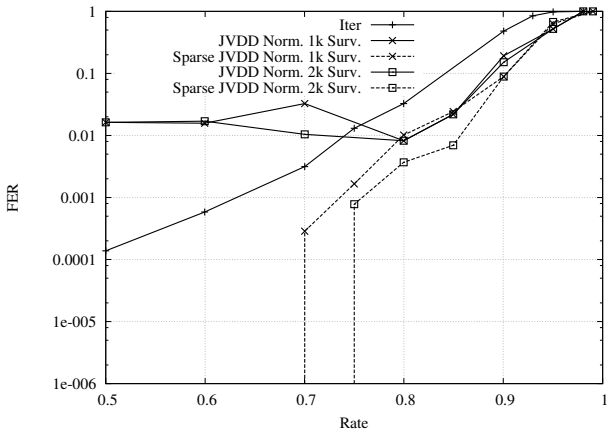


Figure 9. FER performance at different rates at an operating SNR of 6 dB with complexity normalized to 1000 and 2000 survivors for JVDD codes

The sparse construction mitigates the error floor and JVDD is found to outperform the iterative detector at all code-rates. This is due to the fact that iterative detector becomes more efficient for increasing codeword lengths. However, at longer codeword lengths, JVDD becomes more computationally complex. The complexity of JVDD can be controlled through the threshold parameter and increasing the threshold results in retaining more number of survivors in the JVDD trellis. Conversely, this lowers the probability of discarding the MMLC through the metric thresholding section of the JVDD and results in enhanced performance. The performance improvement of JVDD with increasing complexity (average number of survivors) is also depicted in Figure 9.

IV. Conclusions

JVDD algorithm is divided into two stages and the second stage (parity checking) substantially reduces the decoding complexity at low code-rates owing to the increased number of parity checks being performed. However, performance degradation is observed for JVDD at these rates in the low SNR region due an artifact of traditional JVDD code construction. This aspect is analyzed in this paper and attributed to the JVDD code structure where a constant number of ones are placed to the left of the main diagonal. This results in a dense region of ones at the top left corner where the entire rows are filled with ones. A modification to the code construction is proposed in this paper by adjusting the number of ones placed in those rows where the row weight of JVDD codes is less than the available row width. It has been found that too sparse and highly dense region of ones will result in increased JVDD complexity and 50% sparseness will result in optimal JVDD performance. The analysis performed in this paper has been verified through simulation studies.

References

[1] D. Burshtein and G. Miller, “Bounds on the performance of belief propagation decoding,” IEEE Trans. Inf. Theory, vol. 48, no. 1, Jan. 2002, pp. 112–122.

[2] C.-C. Chao, R. McEliece, L. Swanson, and E. R. Rodemich, “Performance of binary block codes at low signal-to-noise ratios,” IEEE Trans. Inf. Theory, vol. 38, no. 6, Nov. 1992, pp. 1677–1687.

[3] S. Haykin, Ed., Communication Systems. 4th Ed., J. Wiley, 2001.

[4] A. Khandekar and R. J. McEliece, “On the complexity of reliable communication on the erasure channel,” in Intl. Symp. Inform. Theory (ISIT), 2001, pp. 7803–7123.

[5] A. J. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” IEEE Trans. Inf. Theory, vol. 13, no. 2, Apr. 1967, pp. 260–269.

[6] S. G. Wilson, Ed., Digital Modulation and Coding. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[7] J. Forney, G., “Generalized minimum distance decoding,” IEEE Trans. Inf. Theory, vol. 12, no. 2, Apr. 1966, pp. 125–131.

[8] D. Chase, “Class of algorithms for decoding block codes with channel measurement information,” IEEE Trans. Inf. Theory, vol. 18, no. 1, Jan. 1972, pp. 170–182.

[9] M. P. C. Fossorier and S. Lin, “Soft-decision decoding of linear block codes based on ordered statistics,” IEEE Trans. Inf. Theory, vol. 41, no. 5, Sep. 1995, pp. 1379–1396.

[10] M. P. C. Fossorier, “Reliability-based soft-decision decoding with iterative information set reduction,” IEEE Trans. Inf. Theory, vol. 48, no. 12, Dec. 2002, pp. 3101–3106.

[11] K. S. Chan, S. S. B. Shafiee, E. M. Rachid, and Y. L. Guan, “Optimal joint viterbi detector decoder (JVDD) over AWGN/ISI channel,” in Intl. Conf. Computing, Networking and Communications (ICNC), Feb. 2014, pp. 282–286.

[12] K. S. Chan and S. S. B. Shafiee, “A study on variable gradient gaussian distribution linear diagonal codes for the JVDD,” in Intl. Conf. Inform. Theory (ICIT), Mar. 2015, pp. 797–801.

 ISBN: 978-1-61208-411-4

Correlation Characteristics of 2-Dimensional Antenna Array Signals in a Multi-Cell Environment

Yoonsu Kim, Wonjin Sung

Department of Electronic Engineering
Sogang University
Email: wsung@sogang.ac.kr

Jonghyun Park

Advanced Standard R&D Lab
LG Electronics
Email: jonghyun10.park@lge.com

***Abstract*—In this paper, we analyze the 3-dimension spatial channel model (3D-SCM) over time and spatial domains. SCM is a realistic channel model for multiple-input multiple-output (MIMO) wireless communication systems since it is known to adequately represent the channel properties with respect to the antenna geometry and conditions of user equipments (UEs). We observe its autocorrelation properties over time and spatial domains as functions of the angle spread and the effective Doppler frequency. The results can be used for efficient design of transmission methods, such as beamforming and precoding using codebooks to fully utilize the channel characterstics of two-dimensional antenna arrays.**

Keywords—**Correlation; MIMO; 2D Arrays; SCM; Angle spread; Doppler frequency.**

I. Introduction

To further increase the bandwidth efficiency of mobile wireless systems, MIMO systems which use multiple antenna elements are being utilized in LTE Release 8 and later releases [1]. It is expected that the number of antenna elements will keep increasing for even higher data rates and bandwidth efficiency. Thus it is important that we understand channel characteristics for different arrangements of antenna elements [2]. 3D-SCM is a channel model for MIMO transmission adopted by the 3GPP Spatial Channel Ad-hoc Group. The model contains a set of parameters, verification methods, and the minimum requirements. It is based on the ray-tracing method and probability models for geometric environments. Various parameters are defined for each of transmission scenarios [3]-[5]. Although these standardization documents specify the setup, implementation procedures, and key characteristics of resulting channels, related studies on how to utilize these results to efficiently design the transmission schemes are not currently well known. More recently, enhancements using full-dimensional MIMO have been discussed and related issues are summarized in [6]. Channel state information reference signaling (CSI-RS) and its enhancement schemes are published as 3GPP meeting documents [7],[8].

In this work, we analyze the 3D-SCM which is implemented by the simulation tool to represent the correlation properties over time and spatial domains. In time domain, we observe autocorrelation of channels for different effective Doppler frequencies of UEs based on their mobility and directions. We compare the autocorrelation of SCM to that generated by the Jake's channel model which follows the zeroth order Bessel function. In spatial domain, we observe the correlation for vertical and horizontal antenna elements respectively, with respect to antenna spacing. We utilize the angle spread to represent the measure of how much beam separation among antenna elements exist [9].

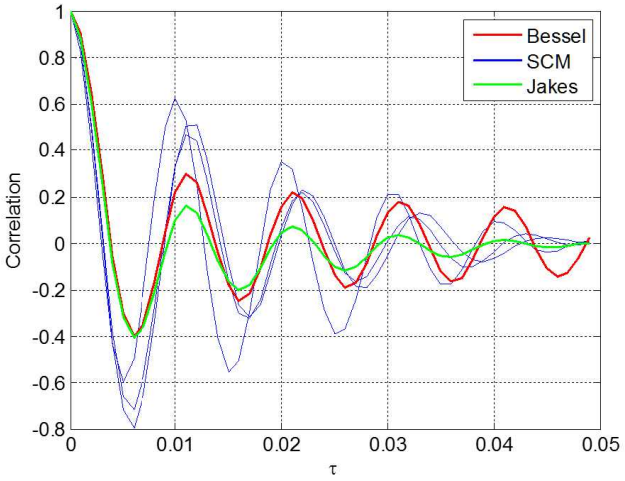
The rest of the paper is organized as follows. In Section II, system model and parameters used for the channel characteristic evaluation are described. Specific results on correlation behavior of the channel model are given in Section III, followed by discussion and conclusions in Section IV.

II. System Model and Parameters

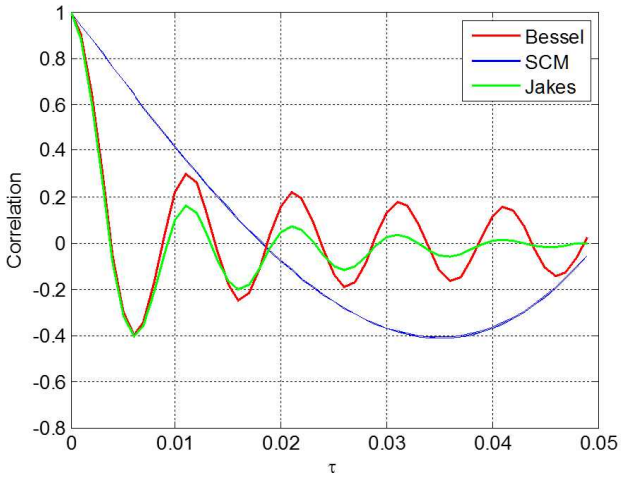
The system model with 19 cells for which each sector has a hexagonal coverage is used, and UEs are assumed to be uniformly distributed in each sector. Each base station (BS) has 16 antenna elements made up of 4 vertical and 4 horizontal antenna arrays, whereas each UE has a single antenna element. Detailed steps of generating 3D-SCM and system parameters used for simulation are similar to those described in [5]. First, environments and parameters are given for UMa (Urban Macro) model, and indoor/outdoor distribution ratio is set as 80 to 20%. Second, departure/arrival angles as well as the pathloss values are determined. Third, delay and power profiles for each cluster are calculated, followed by random coupling among subpaths in each cluster. The fourth step is the assignment of phase values and crosspol factors. Finally, channel coefficients are computed for SCM.

Jakes' channel model is generated for comparison purposes using the Doppler frequency f_d determined as $f_d = v/\lambda$ where v is the UE mobility and λ is the signal wavelength [10]. The zeroth order Bessel function $J_0(\cdot)$ is used for the autocorrelation function (ACF), written as $R[n] = J_0(2\pi f_d T|n|)$ for the time domain [11]. The ACF for the spatial domain can be similarly described as $R[m_H] = J_0(2\pi\delta_H|m_H|d_H/\lambda)$ and $R[m_V] = J_0(2\pi\delta_V|m_V|d_V/\lambda)$ where m_H, m_V are antenna element indices, δ_H, δ_V are angle spreads, and d_H, d_V are antenna spacing. Subscripts H and V represent the horizontal array and vertical array, respectively. We define also κ as $\kappa_H = \delta_H d_H/\lambda$ and $\kappa_V = \delta_V d_V/\lambda$. Doppler frequency value

 ISBN: 978-1-61208-411-4

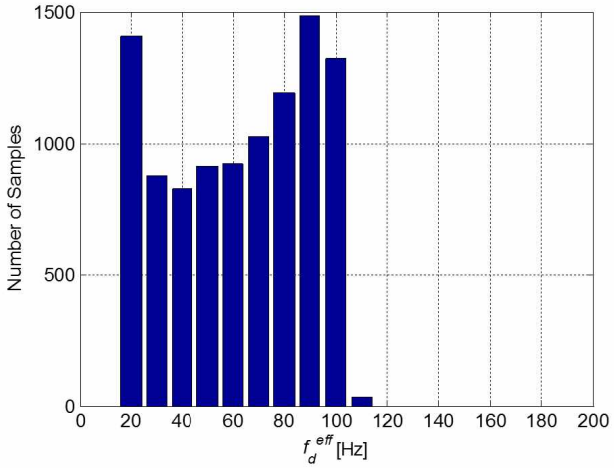


(a)

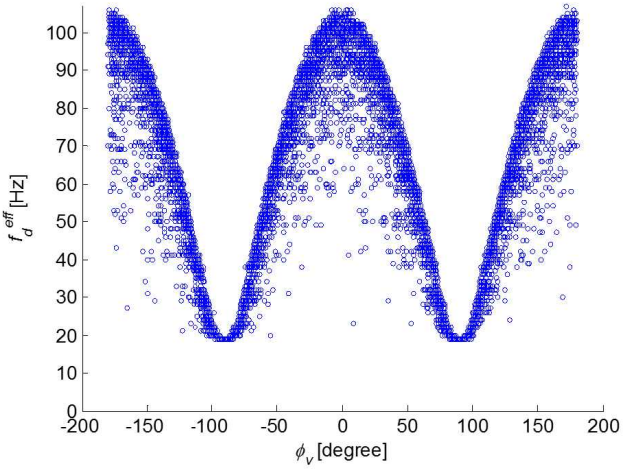


(b)

Figure 1. Comparison of ACFs: (a) $\phi_v = 0^o$ and (b) $\phi_v = 90^o$.



(a)



(b)

Figure 2. Distributions for the effective Doppler frequency: (a) Histogram and (b) f_d^{eff} versus ϕ_{LOS}.

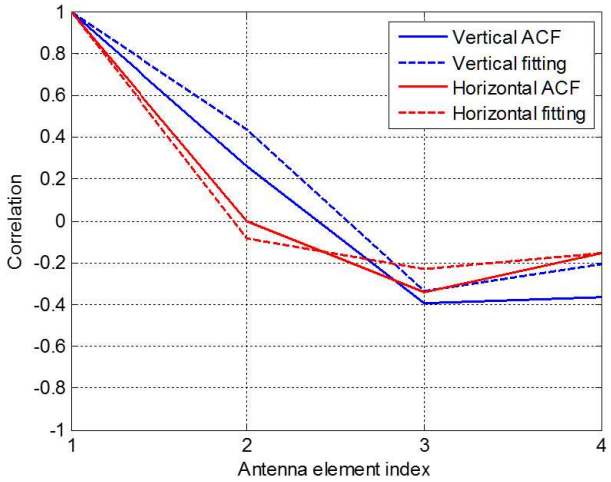
of $f_d = 100$Hz is used for simulation. Effective Doppler frequency is determined as $f_d^{eff} = v \cos \phi_v / \lambda$ which considers the UEs' direction angle ϕ_v. Also, κ^{eff} represents the effective angle spread of each UE.

III. Correlation Characteristics

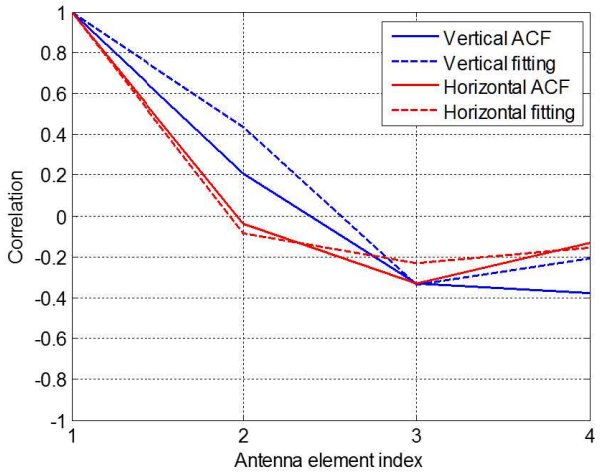
The ACFs as functions of time difference τ are shown in Figure 1. The autocorrelation values were observed by the Monte-Carlo simulation; Channel coefficients are repeatedly generated in random locations within the sector at different time instances, then correlation values at different time intervals are measured by taking the average of those correlation values. Angles $\phi_v = 0^o$ and 90^o represent different moving directions of UEs, which are respectively the same as and perpendicular to the antenna boresight. For UEs with $\phi_v = 0^o$, AFCs are in good agreement to the existing Bessel function model. On the other hand, the UEs with $\phi_v = 90^o$ have high correlation for small time difference values. The distribution of f_d^{eff} and f_d^{eff} versus ϕ_v are summarized in Figure 2.

To verify the ACF characteristics of spatial domain, we first observe the autocorrelation of UEs with respect to vertical and horizontal antenna elements. Then we determine κ_V^{eff} and κ_H^{eff} by the curve fitting method, considering four vertical and horizontal antenna elements for cases of $d_H = d_V = 0.5\lambda$ and $d_H = d_V = 2\lambda$. The purpose of these observations is (1) to verify the accuracy of the curve fitting for horizontal and vertical AFCs, (2) to present the amount of correlation as the spacing between antenna element increases, and (3) to compare the correlation behavior of horizontal and vertical arrays, to better understand the statistical characteristics of SCM. The observation results for these two cases are given in Figures 3 and 4, respectively. Simulations and figure drawings have been produced by Matlab software. The ACFs evaluated in Figure 1 shows that the SCM model implemented produces a reasonable autocorrelation properties needed to simulate the mobile users, as can be seen from the comparison to results produced using the Bessel function and Jakes' fading

 ISBN: 978-1-61208-411-4

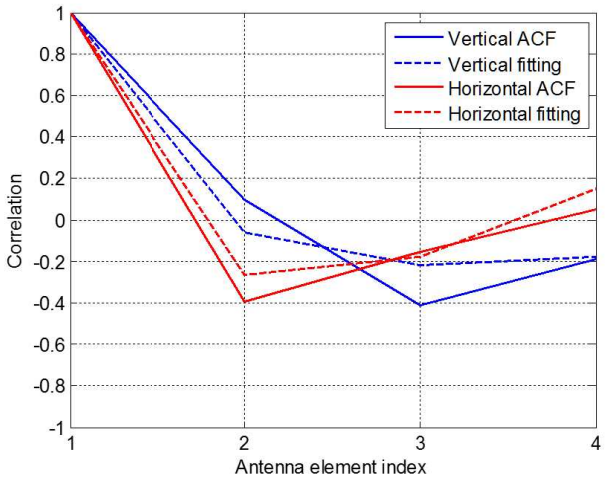


(a)

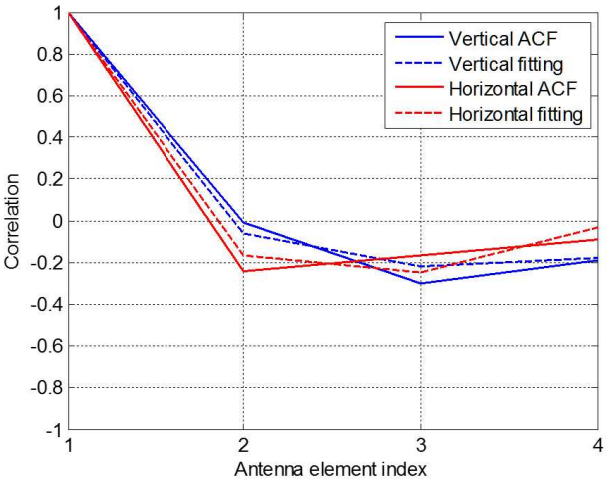


(b)

Figure 3. Actual ACF and curve fitting result for $d_H = d_V = 0.5\lambda$: (a) Specific location and (b) the sector average.



(a)



(b)

Figure 4. Actual ACF and curve fitting result for $d_H = d_V = 2.0\lambda$: (a) Specific location and (b) the sector average.

generation [12],[13]. For the users randomly located over the sector, the distribution for the effective Doppler frequency and the angle-dependent values are given in Figure 2. These results provide a guidance as to which Doppler values to be used for transceiver signal processing.

IV. Discussion and Conclusion

We implemented the 3D-SCM as a realistic correlated channel for MIMO transmission, and analyzed it in both time and spatial domains. We observed the autocorrelation characteristics by experimental results and compared them to existing models. We also determined the effective Doppler frequency in time domain and the effective angle spread in spatial domain by curve fitting. Even when the Doppler frequency of UEs are the same, the autocorrelation function are not the same due to the multipath effects of UEs moving in different directions. By using the distribution results for the effective angle spreads of UEs vertically and horizontally, we observe that the effective angle spreads are mainly dependent on antenna spacing. We can utilize the statistical results of this work for several different purposes of designing the MIMO transmission strategies. For example, theses can be used for channel interpolation, as well as the channel prediction in both domains when only partial knowledge of the actual channel is given. The results can also be applied in selecting beam-forming strategies for the multi-antenna system. Distributions of the channel can be exploited in designing appropriate codebooks to be used in precoding the transmission signal. The estimated Doppler frequency can be used to obtain the precoding vectors in subsequent transmission frames when the exact and full channel information is not present at those frames, by performing extrapolation of previous channel report results.

 ISBN: 978-1-61208-411-4

ACKNOWLEDGMENT

This work was supported in part by the MSIP (Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2014 (B0101-14-0171) and in part by LG Electronics.

REFERENCES

[1] E. Dahlman, S. Parkvall, and J. Sköld, LTE/LTE-Advanced for Mobile Broadband. Academic Press, 2011.
[2] S. Sesia, I. Toufik, and M. Baker, LTE-The UMTS Long Term Evolution: From theory to practice, 2nd ed. Chichester, U.K.: Wiley, 2011.
[3] 3GPP TS 36.814, v9.0.0, Evolved Universal Terrestrial Radio Access (E-UTRA); Further advacements for E-UTRA physical layer aspects (Release 9), Mar. 2010.
[4] 3GPP TR 25.996, v11.0.0, Technical Specification Group Radio Access Network; Spatial channel model for MUltiple Input Multiple Output (MIMO) simulations (Release 11), Sept. 2012.
[5] 3GPP TR 36.873, v1.2.0, Technical Specification Group Radio Access Nework; Study on 3D channel model for LTE (Release 12), Sept. 2013.
[6] 3GPP TR 36.897, v1.0.0, Study on Elevation Beamforming/Full-Dimension (FD) MIMO for LTE (Release 13), May. 2015.
[7] 3GPP TSG RAN WG1 #81 Meeting Document R1-153596, WF on CSI-RS enhancements, Fukuoka, Japan, 25-29 May 2015.
[8] 3GPP TSG RAN WG1 #81 Meeting Document R1-152894, Discussion on spatially sub-sampled non-precoded CSI-RS, Fukuoka, Japan, 25-29 May 2015.
[9] A. Abdi and M. Kaveh, "A space-time correlation model for multielement antenna systems in mobile fading channels," IEEE J. Select. Areas in Commun., vol. 20, no. 3, Apr. 2002, pp. 550–560.
[10] Y. R. Zheng and C. Xiao, "Improved models for the generation of multiple uncorrelated Rayleigh fading waveforms," IEEE Commun. Lett., vol. 6, no. 6, July 2002, pp. 256–258.
[11] K. E. Baddour and N. C. Beaulieu, "Autogressive modeling for fading channel simulation," IEEE Trans. Wireless Commun., vol. 4, no. 4, July 2005, pp. 1650–1662.
[12] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, Time Series Analysis: Forecasting and Control, 3rd ed. New Jersey, U.S.A: Prentice Hall, 1994.
[13] S. K. Lucas and H. A. Stone, Evaluating infinite integrals involving Bessel functions of arbitrary order, J. Computational and Applied Math., vol. 64, no. 3, Dec. 1995.

 ISBN: 978-1-61208-411-4

Securing Commercial Ad Broadcasting in Vehicular Ad Hoc Networks

Kevin Daimi
Computer Science and Software Engineering
University of Detroit Mercy
Detroit, USA
daimikj@udmercy.edu

Mustafa Saed and Scott Bone
HATCI Electronic Systems Development
Hyundai-Kia America Technical Center
Superior Township, USA
{msaed, sbone}@hatci.com

***Abstract*—Commercial advertising via vehicular ad hoc networks (VANETs) is a promising application. It allows organizations to target drivers and passengers with the aim of promoting their products and services. The implementation of such an application will not be successful without guaranteeing that these ads will not include any malicious information, and the ads will be broadcasted. This paper will apply a cryptographic protocol to secure the dissemination of commercial ads. Secure incentives for drivers reading/watching the ad will be introduced. Cheating, including multiple incentives for the same ad by the same driver will be dealt with.**

Keywords-Commercial Ad; Vehicular Ad Hoc Networks; Security; Security Architecture; Secure Incentives.

I. INTRODUCTION

Vehicular ad hoc networks (VANETs) treat vehicles in their vicinity as wireless nodes. Vehicles within this area can communicate with each other. Any vehicle exiting the zone will lose its communication with that VANET. Vehicular ad hoc networks allow vehicles to broadcast messages to all other vehicles within the range. This presents a great opportunity for various applications to be implemented on vehicles using their computing power and storage capabilities. Safety-critical information including speed, heading, and position in addition to various warning on accidents and road conditions, and infotainment can be broadcasted by vehicles using vehicle-to-vehicle communication. Vehicular ad hoc networks (VANETs) are a subclass of mobile ad hoc networks (MANETs). However, VANETs present characteristics that are noticeably different from many generic MANETs. VANETs are considered as a pledging style for future Intelligent Transportation System (ITS). They possess no static infrastructure. Consequently, they expect vehicles to deliver network functionality [1]. VANETs are extricated from other classes of ad hoc networks through their hybrid network architectures, node movement features, and non-traditional new application settings. As a result, VANETs call for numerous unique research challenges. Furthermore, the design of an effective routing protocol for VANETs is undoubtedly vital [2]. Vehicular ad hoc networks can provide a wide variety of services. However, they are subject to a number of challenges including network architecture, protocols for physical and link layers, and routing algorithms [3]. Vehicular ad hoc networks will not only make safety and lifesaving applications a reality, but will also turn out to be a formidable communication instrument for their users [4].

With the increasing number of various attacks on wireless networks, security becomes a critical challenge for VANETs and will continue to be so even after it is widely implemented. VANETs are subject to many attacks including denial of service, Sybil, hardware, software, sinkhole, impersonation, and flooding attacks. To ensure effective security, the security requirements; availability, authentication, integrity, confidentiality, and non-repudiation must be satisfied.

Considerable work on VANETs security has been pursued to ensure the above mentioned security requirements are met. Most of this work adopted cryptology. Symmetric and, asymmetric cryptology, and tamper resistance hardware were suggested. For some authors, cryptographic certificates were an option. Other authors investigated various threats, particularly threats related to security requirements, and created various security protocols. Standardization related to approaches of furnishing security services and protecting driver's privacy were analyzed. Gillani et al. [5] examined several aspects of VANETs security including security threats, challenges in providing security in VANETs environment, security requirements, and attributes of security solutions. The need for robust VANET networks is obviously related to their security and privacy characteristics. Various types of security problems and challenges of VANET, and a set of solutions to solve these challenges and problems have been analyzed and discussed in [6]. Al-Kahtani [7] stressed that designing security mechanisms to authenticate and validate transmitted messages between vehicles, and remove intruders from the network are substantially critical in VANETs. The author also reported several existing and possible security attacks and techniques to enhance the security of VANETs. Security and privacy are obligatory in vehicular communications for successful acknowledgment and utilization of VANET technology. Every vehicular application must be meticulously tested for security before it is

 ISBN: 978-1-61208-411-4

implemented in the real world. Simulation tools have proved to be very effective for such testing [8]. The security of VANET has mostly inspired the current research efforts. Thorough solutions to safeguard the vehicular ad hoc networks against adversaries and attackers still need to be developed to arrive at an adequate level for both the driver and manufacturer to achieve safety of drivers and security of applications and infotainment [9]. Details of further attempts to secure VANETs and its applications could be found in [10]-[17].

Commercial advertising via vehicular ad hoc networks (VANETs) is a promising application. It allows organizations to target drivers and passengers with the aim of promoting their products and services that can possibly solve a problem in their lives. Advertising through VANETs will get the word out rapidly and more visibly to customers. Customers in a vehicle would have the opportunity to listen to or watch ads that serve their needs. Such advertising can compete with TV ads due to the fact that many audiences ignore most of the ad breaks on TV or get/do something else during those breaks.

The growth in market prospects and potentials necessitates further research on mobile marketing, such as mobile advertisement. Mobile advertisements intermingle with customers on one-to-one basis via messages through the use of mobile devices [18]. Wireless technology has initiated new channels of marketing communication and innovation of advertisement media, such as the mobile advertisement platform. Mobile advertisement relies on the use of wireless networks to dispense information about products to consumers in a localized, specialized and customized manner [19]. People making use of modern wireless technology are more likely to consider mobile networks as their daily entertainment device than watching TV and possibly reading newspapers [20], [21]. Advertisement with mobile networks can highly target customers who will find reading or watching advertisement through mobile network more enjoyable and valuable [22]. These reasons provide mobile marketing with an effective means for advertisers to directly reach out for their potential consumers more effectively [23].

A secure incentive framework for commercial ad dissemination in VANETs was introduced by Li et al. [24]. The presented approach relied on public key infrastructure to provide secure incentives for cooperating nodes. The framework relied on vehicles receiving ads and disseminating them to other vehicles. The possibility of cheating by some drivers who can send receipts without even examining the ad is very high. Multiple receipts for the same ad by the same vehicle will go undetected. Furthermore, the authors used public keys to encrypt ads. Public keys are inefficient for encrypting large messages. Zhu [25] introduced the security requirements for service-oriented vehicular networks. Commercial content distribution is one of these services. Secure payments are possibly needed for some commercial application in VANETs [26].

This paper proposes a secure commercial ad broadcasting via VANETs. The security architecture for the dissemination of the ad is integrated with the vehicular ad hoc network security architecture proposed by the authors in [27]. Various cryptology protocols will be presented, and the security of incentives will be implemented. The approach followed in this paper also treats possible cheating including drivers passing the ad code to their friends to claim incentives without reading/watching the ad, and requesting multiple incentives for the same ad by the same vehicle. Section II presents the ad broadcasting architecture. Section III demonstrates public key certificates distribution. Organization to ad administration communication and state-level RSU to ad administration communication are introduced in Sections IV and V respectively. Section VI provides the state-level RSU to county-level RSU communication. Sections VII, VIII, and IX describe county-level RSU to city-level RSU communication, city-level RSU to street-level RSU communication, and street-level RSU to vehicle communication respectively. The paper is concluded in Section X.

II. AD BROADCASTING SECURITY ARCHITECTURE

The ad broadcasting architecture is superimposed on the multi-level security architecture for vehicular ad hoc networks introduced by the authors in [27]. It is re-drawn to serve the purpose of the ad broadcasting security architecture. The ad issuing organization (AORG), ad authority (AUTH), and the ad administration authority (ADMN) are added to it. Fig. 1 illustrates the ad broadcasting security architecture that will guide the security protocols. The right hand side of this figure represents the security architecture for vehicular ad hoc networks mentioned above. This is augmented by the left hand side part to include secure ad dissemination. Note that apart from the box for RSU_C, there supposed to be a number of boxes for all levels on the right hand side of Fig. 1 to indicate many states, counties, and cities.

The roadside units (RSU_S) are organized in a hierarchal fashion. The root of this tree is the Country-Level RSU (RSU_C). State-Level RSUs (RSU_S) are connected to the (RSU_C). Likewise, County-Level RSUs (RSU_{CO}), City-Level RSUs (RSU_{CI}), and Street-Level roadside units (RSU_{ST}) are connected to RSU_S, RSU_{CO}, and (RSU_{CI}) respectively. Each Street-Level RSU is in charge of all vehicles passing through the street (or portion of the street for long streets) under its authority. RSUs within the same level can only communicate through the parent node they belong to. The computing power and capacity of RSU increases when moving upwards through the tree. Detailed information about vehicles is stored at the State-Level RSU (RSU_S). With the exception of RSU_C, there are many RSU_S, RSU_{CO}, RSU_{CI}, and RSU_{ST} at their levels. However, only one RSU of each is shown in Fig. 1.

The ad authority (AUTH) is in charge of issuing certificates to the ad issuing organizations (companies interested in promoting their products or services), the ad administration authority (ADMN), and the State-Level RSU (RSU_S). For each state, there is only one ad authority and one ad administration authority. In other words, one AUTH and one ADMIN will manage ads for the cities within the state.

 ISBN: 978-1-61208-411-4

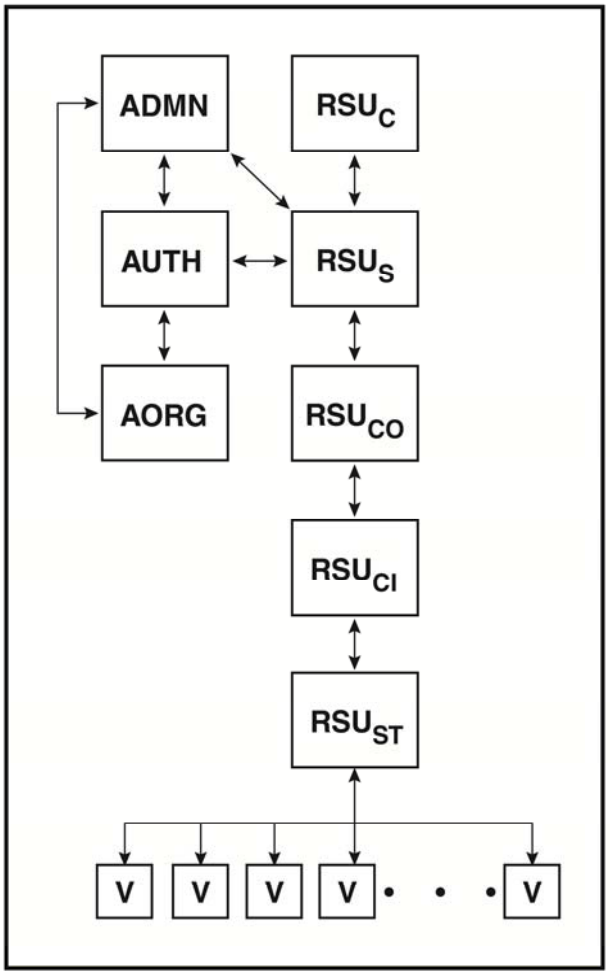


Figure 1. Ad broadcasting security architecture

The communication between RSU_{CT} and RSU_{ST}, and RSU_{ST} and vehicles follows the security protocols used in [27]. In these protocols, RSUs at each level create their own private and public keys and exchange their public keys. Each parent RSU creates a master key. The master key and the ID of the parent node are encrypted with the public keys of the children nodes and forwarded to them. At this point all nodes have a shared master key with their parent nodes. The public and private keys are then discarded. Each parent node creates a session key and encrypts it with the master key. Once the session keys are obtained, messages can be exchanged. To send any message between a child and a parent, the hash function of the message is appended to the message and both are encrypted with the session key. Note that Street-Level RSU (RSU_{ST}) creates public key certificates for all vehicles entering its street. This will be used by vehicles when broadcasting messages to other vehicles.

The ad material is sent by the ad issuing organizations (AORG) to the ad administration authority (ADMN) together with the ad ID (AID) and ad period (ADP). ADMN checks the ad against the legal, social, and ethical constraints. It later negotiates the cost with the ad issuing company. The cost includes what the administration authority charges, the State-Level RSU charges, and the coupon amount/number of points that will be allocated to vehicles reading or watching the ad. The State-Level charges include the amounts allocated to itself, the county, city, and Street-Level RSUs. Every ad contains an ad code (ADC), which will be used for incentives purposes. Upon completion of this part of the protocol, the ad material is forwarded by ADMN to the State-Level (RSUs). The Street-Level RSUs will receive the ad material from its parent City-Level node (RSU_{CI}) and securely broadcast it to the vehicles within its street authority. Ads (clip or text and images) are mainly large messages. Public key cryptology tends to be very slow and inefficient when dealing with such large messages. Therefore, the ad administration authority (ADMN) will create two session keys, K_{S1} and K_{S2}, which will be shared with AORG and RSU_S respectively. These two keys will be valid until the ad is completely administered. This will occur when ADMN receives the charges from AORG.

The Street-Level RSUs (RSU_{ST}) ensures via secure communication that vehicles within its responsibility have read/watched the ad when they receive the ad code (ADC). This can only be obtained when reaching the end of the ad. Information regarding the participating vehicles will be forwarded to the parent RSU_{CI} for charging purposes. At the expiration date of the ad, the City-Level RSU will send info about all the participating vehicles to the RSU_S via the RSU_{CO}. The RSU_S will send the charging information for all the cities within that state to the ad administration authority (ADMN) for charging purposes. Any incentive system for participating vehicles can be utilized provided it is secure. However, in this paper, coupon and points redemption will be used. RSU_S, RSU_{CO}, RSU_{CT} and RSU_{ST} will receive dollar amounts.

To better understand the protocols, the participating parties are introduced in Table 1. Table 2 depicts the protocols notations and symbols used in the protocols.

III. PUBLIC KEY CERTIFICATES DISTRIBUTION

The Ad Administration Authority (ADMN), the Ad Issuing Organizations (AORG), and the State-Level RSU (RSU_S) request their public key certificates from the Ad Authority (AUTH). The public key of the Ad Authority, PU_{AUTH} is made known to all these parties above. The public key, ID, and a nonce for each party are encrypted with the public key of the Ad Authority and forwarded to it.

$RSU_S \rightarrow$ AUTH: $E[PU_{AUTH}, (PU_S \| ID_S \| N_S)]$
ADMN $\rightarrow$ AUTH: $E[PU_{AUTH}, (PU_{ADMN} \| ID_{ADMN} \| N_{ADMN})]$
AORG $\rightarrow$ AUTH: $E[PU_{AUTH}, (PU_{AORG} \| ID_{AORG} \| N_{AORG})]$

The Ad Authority decrypts each message and creates public key certificates for the three parties and attaches the original nonce encrypted with the party's public key.

$CR_S = E[PR_{AUTH}, (PU_S \| ID_{US} \| T_1 \| T_2)]$
$CR_{ADMN} = E[PR_{AUTH}, (PU_{ADMN} \| ID_{ADMN} \| T_1 \| T_2)]$
$CR_{AORG} = E[PR_{AUTH}, (PU_{AORG} \| ID_{AORG} \| T_1 \| T_2)]$

 ISBN: 978-1-61208-411-4

AUTH → RSU_S: CR_S || $E(PU_S, N_S)$
AUTH → ADMN: CR_{ADMN} || $E(PU_{ADMN}, N_{ADMN})$
AUTH → AORG: CR_{AORG} || $E(PU_{AORG}, N_{AORG})$

The certificates include a timestamp, T_1, and a certificate validity period (expiration date), T_2. The original nonce are encrypted with the public key of the party and attached for further assurance that the message is not a replay.

TABLE 1. PARTICIPATING PARTIES

Symbol	Role
AUTH	Ad Authority
ADMN	Ad Administration Authority
AORG	Ad Issuing Organization
RSU	Road side unit
RSU_C	Country-Level RSU
RSU_S	State-Level RSU
RSU_{CO}	County-Level RSU
RSU_{CI}	City-Level RSU
RSU_{ST}	Street-Level RSU
V	Vehicle

TABLE 2. PROTOCOL NOTATIONS

Symbol	Meaning
PU_C , PR_C	Public & private key of Country-Level RSU
PU_S , PR_S	Public & private key of State-Level RSU
PU_{CO} , PR_{CO}	Public & private key of County-Level RSU
PU_{CI} , PR_{CI}	Public & private key of City-Level RSU
PU_{ST}, PR_{ST}	Public & private key of Street-Level RSU
PU_V, PR_V	Public & private key of vehicle
K_M, K_S	Symmetric Master and session Keys
K_{MS}, K_{SS}	K_M, K_S shared by state and county RSUs
K_{MCO}, K_{SCO}	K_M, K_S shared by county and city RSUs
K_{MCI}, K_{SCI}	K_M, K_S shared by city and street RSUs
||	Concatenation
E	Encrypt
→	Send to
H(M)	Hash of message M
T_1	Issue time
T_2	Expiration time
AID	Ad ID
ADP	Ad Period
ADC	Ad Code
C/P	Coupon amount/Number of points
ID_V, ID_{VA}	Real and Anonymous ID of vehicle
ID_S,	ID of State-Level RSU
ID_{CO}	ID of County-Level RSU
ID_{CI}	ID of City-Level RSU
ID_{ST}	ID of Street-Level RSU
ID_S	ID of State-Level RSU
ID_{AORG}	ID of Ad Issuing Organization
ID_{ADMN}	ID of Ad Administration Authority
PU_{AORG}	Public key of AORG
PR_{AORG}	Private key of AORG
PU_{ADMN}	Public key of ADMN
PR_{ADMN}	Private key of ADMN
T_{AORG}	Time stamp added by AROG
N_i	Nonce, i = S, AROG, ADMN
K_{S1}	Session key shared by ADMN and AORG
K_{S2}	Session key shared by ADMN and RSU_S

IV. ORGANIZATION-TO-AD ADMINISTRATION COMMUNICATION

This sub-protocol involves the ad approval and costing, and securely handling incentives.

A. Ad Approval and Costing

During this communication, the ad will be either accepted or rejected. In addition, the charges will be set. These charges will include the incentives which will be paid to vehicles. These will be taken care of later in this paper. The integrity of all messages is important.

The Ad Issuing Organization (AORG) and Ad Administration Authority (ADMIN) exchange certificates, validate the currency of each other's certificate, and extract the public key and ID of the other party. ADMN creates a session key, K_{S1}, to be shared with AORG. This session key and the ID of ADMN are encrypted with PR_{ADMN} and then with PU_{AORG} and forwarded to AORG. After carrying out the needed decryptions to get K_{S1} and verifying the sender, AORG sends a request to ADMN for ad dissemination. The request includes the ad ID (AID), the ad as a clip or text (AD), ad period (ADP), hash of the ad, H(AD), AORG's ID, and a timestamp, T_{AORG}. AD and H(AD) are encrypted with K_{S1}. The rest are encrypted with AORG's private key and then with the public key of ADMN.

$Z = E[K_{S1}, AD || H(AD)]$
$X = AID || ID_{AORG} || ADP || T_{AORG}$
AORG → ADMN: $E[PU_{ADMN}, E(PR_{AORG}, X)] || Z$

ADMN will first decrypt the first part of the message using its private key and then with the public key of AORG. It then decrypt Z with K_{S1} to get AD and H(AD), calculates the hash code of AD and compare it with H(AD). It will also check the timestamp to ensure the message's currency. Having verified the hash and timestamp, ADMN will examine the Ad to see if is not violating any legal, social, or ethical requirements. It then, extracts the Ad Code (ADC), which can only be obtained when the end of the ad is reached. The ADC will be used for incentive purposes in the future. ADMN also uses it as an assurance to AORG that the ad has been processed by ADMN. Finally, a message containing AID, ADC, ID, Reject/Accept (R/A), and Ad Dissemination Cost (ACOST) will be encrypted with AORG's public key PU_{AORG}. This implies one of the messages below will be sent depending on whether the ad is accepted or rejected. Let $Y = AID || ADC || ID_{AORG} || ID_{ADMN}$.

ADMN → AORG: $E[PU_{AORG}, A || Y || ACOST]$ or
ADMN → AORG: $E[PU_{AORG}, R || Y]$

If the ad is rejected, no further communication for that ad will be followed. Otherwise, AORG decrypts the message and verifies AID and ADC. It either agrees or disagrees with the cost. ID_{AORG} and ID_{ADMN} are used as assurance components. If AORG agrees, it sends the following message to ADMN:

 ISBN: 978-1-61208-411-4

AORG → ADMN: E[PU_{ADMN}, (PR_{AORG}, Y || ACOST || AGREE)]

Upon receiving this message and decrypting it, ADMN will verify the agreement and the stated cost. Once again, the two IDs, ID_{AORG} and ID_{ADMN}, are used for assurance purposes.

B. Secure Incentives Handling

ADMN adds the total amounts for the coupons/points received from the RSU_S to its charges and the charges of the state. This represents the total amount charged for that ad.

M = ID_{AORG} || ID_{ADMN} || ADC || AID || TOTAL
ADMN → AORG: E[PU_{AORG}, E(PR_{ADMN}, M || H(M)]

AORG will subtract ACOST from TOTAL to get the total incentives for vehicles. It will then divide the result by the coupon value or number of points allocated to this ad to find out how many vehicles read/watched the ad. Having done that, TOTAL will be transferred to ADMN using any secure approach.

V. State-Level RSU-to-AD Administration Communication

This section introduces the ad material and incentive forwarding sub-protocols.

A. Ad Material Forwarding

The Ad Administration Authority forwards the ad material to the State-Level RSU. In addition, ADMN sends the monetary amount to the RSU_S. It will either accept the ad or return to ADMN in case of any problem. Both ADMN and RSU_S swap over certificates. If the certificate is valid, the ID and public keys are retrieved. ADMN creates a session key, K_{S2}, to be shared with RSU_S. ADMN then forms a message containing ID_{ADMN}, the ad ID (AID), ad code (ADC), ad period (ADP), ID_{AORG}, and coupon amount or number of points (C/P) all encrypted with PR_{ADMN} first and then with PU_S. It then attaches AD || H(AD) after encrypting them with K_{S2}. The resulting message is sent to RSU_S.

Z = E[K_{S2}, AD || H(AD)]
X = ID_{ADMN} || AID || ADC || ADP || ID_{AORG} || C/P
ADMN → RSU_S: E[PU_S, E(PR_{ADMN}, X)] || Z

RSU_S performs the needed decryptions, verifies the hash code of the ad equals H(AD), ensures the ad code is the same as ADC and ID_{ADMN} is a valid ID. It also validates the ad period to make sure it is not an expired ad. If there is an issue with all these checks, a message containing the problem will be sent. Examples include "Invalid ID" and "Mismatched ADCs." The world "PROBLEM" will be used. If there is no problem, "VALID" will be attached to the message.

Y = ID_S || ADC || AID
RSU_S → ADMN: E(PU_{ADMN}, Y || PROBLEM)
RSU_S → ADMN: E(PU_{ADMN}, Y || VALID)

B. Incentive Forwarding

The responsibility of the RSU_S in this communication is to forward a list of vehicles to the ADMN for incentives purposes. The information about vehicles involved in the ad will be received from the County-Level RSUs (RSU_{CO}).

At the expiration date of an ad (ADP), the RSU_S first ensures that no vehicle within the state will get multiple incentives for the same ad. Having done that, the State-Level RSU sends a message, M, containing the name of the driver, ID_V, address (ADR), ADC, AID, C/P, and H(M) encrypted with PR_S and then with ADMN's public key.

M= ID_S || NAME || ADR || ADC || ID_V || AID || ADC || C/P
RSU_S → ADMN: E[PU_{ADMN}, E(PR_S, M || H(M))]

ADMIN decrypts this message and verifies there are no duplicate incentives for the same ad for the ID_V that was received. The total incentives (coupon amount or number of points) are then updated. This will be done for all the different ads. At the end of the month, a coupon or total number of points will be mailed to the vehicle's driver.

VI. State-Level RSU-to-County-Level RSU Communication

The State-Level RSU maintains vehicle database. It transmits the ad materials to all counties and receives all the anonymous IDs used for each vehicle at all counties, and the incentive details for all vehicles. It uses the received information to update its database of vehicles. In US, the minimum number of counties is 3 and the maximum is 254. Large counties will have more streets. This will demand more street-level RSUs (RSU_{ST}) and more advanced equipment's to improve performance. The stretch of a street assigned to an RSU will designate the maximum number of vehicles under the responsibility of that RSU. Therefore, the limit on the number of vehicles is only determined by the capacity of the allocated street section.

The state forwards the ad material continued in M below after encrypting it with the session key, K_{SC}, shared with RSU_{CO} to the County-Level RSU:

M= ID_{CO} || ID_S || AID || ADC || AD || H(AD) || ADP || C/P
RSU_S → RST_{CO}: E(K_{SC}, M|| H(M))

The RSU_S receives the ID_V and ID_{VA} for all vehicles from all counties. The records in the State-Level database will be updated for each vehicle. Note, ID_{ST} indicates where the ID_{VA} was issued. In other words, it is the ID of street accommodating the vehicle at that time.

M = ID_{CO} || ID_S || ID_{ST} || ID_V || ID_{VA}
RSU_{CO} → RST_S: E(K_{SC}, M|| H(M))

 ISBN: 978-1-61208-411-4

Each county will send the incentives information for all vehicles within its cities after verifying no duplications exist for a vehicle among its cities with regards to the same ad.

$M = ID_{VA} || ID_{CO} || AID || ADC || ADP || C/P$
$RSU_{CO} \rightarrow RST_S: E(K_{SC}, M|| H(M))$

VII. County-Level RSU-to-City-Level RSU Communication

In this communication sub-protocol, the ad material dispatching, storing vehicle information and incentive handling will be dealt with.

A. Ad Material Dispatching

The RSU_{CO} sends the ad material to the City-Level RSUs in addition to its ID and the ID of each RSU_{CI} within that county.

$M= ID_{CI} || ID_{CO} || AID || ADC || AD || H(AD) || ADP || C/P$
$RSU_{CO} \rightarrow RST_{CI}: E(K_{SCO}, M|| H(M))$

K_{SCO} is the shared session key between RSU_{CO} and RSU_{CI}. RSU_{CI} will decrypt this message, verify the sender, check the integrity of the ad, and obtain the ad material.

B. Storing Vehicle Information

The County-Level RSU (RSU_{CO}) receives all vehicle IDs with all their ID_{VA}'s, and the location where ID was issued. This location is in fact the Street-Level RSU's ID. The RSU_{CI} sends the RSU_{CO} the following information about each vehicle at each location (street):

$M = ID_{CO} || ID_{CT} || ID_{ST} || ID_V || ID_{VA}$
$RSU_{CI} \rightarrow RST_{CO}: E(K_{SCO}, M|| H(M))$

Here, K_{SCO} is the session key shared between RSU_{CI} and RSU_{CO}. There could normally be a number of such messages for the same vehicle, but for different ads. The RSU_{CO} will store this information together with that received from the State-Level RSU as mentioned above in its database. This history information will be beneficial for law enforcement authority to trace a vehicle if a need arises.

C. Incentive Handling

Having verified there are no multiple incentives for the same ad, the RSU_{CI} sends the message $E(K_{SCO}, M|| H(M))$ to the RSU_{CO} at the expiration of the ad.

$M = ID_{CO} || ID_{CI} || ID_V || ID_{VA} || AID || ADC || ADP || C/P$
$RSU_{CI} \rightarrow RST_{CO}: E(K_{SCO}, M|| H(M))$

After decrypting the message, verifying the sender, and validating the ad material introduced in M above, the County-Level RSU checks that there are no multiple vehicle incentive requests by the same vehicle for the same ad among all the cities belonging to that county.

VIII. City-Level RSU-to-Street-Level RSU Communication

The City-Level RSU, RSU_{CI}, receives vehicles IDs and all anonymous vehicle IDs from the Street-Level RSU. It also receives the needed ad information for incentive purposes. RSU_{CI} sends the ad material to all Street-Level RSUs within the city.

A. Vehicles ID Storing

Each RSU_{ST} send a list of real IDs and anonymous IDs for each vehicle passing through that street. As mentioned in Section II, RSUs communicate using a shared session key. Therefore, the list of IDs and the hash code of the list is encrypted with the shared session key for Street-Level and City-Level RSUs (K_{SCI})

$M = ID_{ST} || ID_{CI} || ID_V || ID_{VA}$
$RSU_{ST} \rightarrow RST_{CI}: E(K_{SCI}, M|| H(M))$

The RST_{CI} updates its database to add all new ID_{VA} issued for the vehicle during that period.

B. Sending Ad Material

The RST_{CI} sends a message, M, composed of its ID, the Street-Level ID, AID, ADC, ADP, C/P, and AD. The hash code of M is also attached.

$M = ID_{ST} || ID_{CI} || AID || ADC || ADP || AD || H(AD) || C/P$
$RSU_{CI} \rightarrow RST_{ST}: E(K_{SCI}, M|| H(M))$

The Street-Level RSU confirms the sender and the message integrity. It then saves AID, ADC, ADP, C/P, and AD.

C. Incentive Forwarding

The Street-Level RSU sends its RSU_{CI} incentive messages for each participating vehicle:

$M = ID_{VA} || ID_{ST} || AID || ADC || ADP || C/P$
$RSU_{ST} \rightarrow RST_{CI}: E(K_{SCI}, M|| H(M))$

The City-Level RSU checks that there are no duplications for any ad's incentives within its streets. In other words, because RSU_{CI} has the incentive information from all its streets, it makes sure no vehicle has sent multiple ADC for the same ad whether within the same street (driving through it more than once) or at various streets within the city. At the end, each ad participating vehicle will have just one incentive for an ad. Definitely, multiple incentives for different ads are acceptable.

IX. Street-Level RSU-to-Vehicle Communication

The Street-Level RSU, RSU_{ST}, receives the real ID of the vehicle, ID_V, when entering its zone, and provides its public key, PU_{ST}, to that vehicle. The RSU_{ST} uses a three-measurement technique [6] to create an anonymous ID, ID_{VA}, for the vehicle. Each vehicle will create its own public and

 ISBN: 978-1-61208-411-4

private keys (PU_V, PR_V), and forwards its public key to its RSU_{ST}.

RSU_{ST} creates a secret label, L, for each vehicle entering its zone. It creates a random key, K_L, and encrypts the ad ID (AID) and ID_{VA} with it. In other words, $L = E(K_L, AID \| ID_{VA})$. K_L is not shared with the vehicle. It will only be used once for each ad to control cheating. Without this label, vehicles can cheat by sending the ADC to other vehicles within the street, or another street, possibly in another city. With the absence of such a label, vehicles receiving the ADC can submit the required details without reading/ watching the ad and to earn incentives. L is encrypted with the public key of the vehicle and forwarded to it. Vehicles requesting incentives should attach L to other incentive requirements.

$RSU_{ST} \rightarrow V: E(PU_V, L)$

RSU_{ST} sends the ad materials to the vehicle. It appends the ad ID (AID), ID_{VA}, C/P, the ID of the Street-Level RSU, ID_{ST}, AD, and the hash code of the ad, H(AD) together to get the message X. The hash function is used to ensure the integrity of the ad. RSU_{ST} relies on broadcasting messages. To achieve broadcasting, the RSU_{ST} selects a random key, K_r, to encrypt X. It then encrypts K_r with the public key, PU_V, of each vehicle. Finally, the ID of the vehicle is attached to both encrypted messages and broadcasted to all vehicles in the zone.

$X = ID_{VA} \| AID \| ID_{ST} \| AD \| H(AD) \| C/P$
$RSU_{ST} \rightarrow V: ID_{VA} \| E(PU_V, K_r) \| (K_r, X)$

Recognizing their IDs, vehicles will decrypt with their public key PU_V first to get K_r, and then with K_r to get the message X. The vehicle will verify the sender. It then ensures the message is integral. Later, the vehicle's driver will decide if he/she is interested in the ad based on the value of C/P. To be eligible for incentives, the driver must watch the clip to the end, or read the text of the ad to the end in order to extract the ad code (ADC). The ad code is the proof that will be used for providing incentives. If the ad is followed to the end, the vehicle sends a message containing the ADC, AID, anonymous ID of the vehicle, ID of RSU_{ST}, and the label (L) all encrypted first with the vehicle's private key and then with the public key of the RSU_{ST}.

$V \rightarrow RSU_{ST}: E[PU_{ST}, E(PR_V, AID \| ID_{VA} \| ID_{ST} \| ADC \| L)]$

After carrying out the decryptions and recognizing the sender, the RSU_{ST} verifies the received ADC matches the ADC of one of the ads, and ensures the AID in the message is the same as the AID of that ad. Finally, it verifies the ad period (ADP), which was forwarded to it by the RSU_{CI} to ensure the ad is still valid. If there is any problem, the received message is ignored. Finally, verification against cheating is carried out by decrypting L with K_L and checking that AID and ID_{VA} of the label match the received AID and ID_{VA}. If verification is positive, an acknowledgment (ACK) is sent to the vehicle. The Keys, K_L and K_r, will be discarded once the expiration date of the ad in question is reached.

$RSU_{ST} \rightarrow V: E[PU_V, E(PR_{ST}, AID \| ID_{VA} \| ID_{ST} \| ACK)]$

X. CONCLUSION

The advent of vehicular ad hoc networks (VANETs) widely opened the door for various commercial applications. An important application is the commercial ad broadcasting. For such application to be successful and effective, dissemination of ad should be carried out in a secure manner to protect various communications. Securing the ad dissemination without providing incentives will render the application ineffective as many drivers will just ignore the ads. This paper introduced a secure architecture, which is implemented by a secure protocol to protect communications and incentives. The protocol also prevented dishonest drivers, if any, from cheating.

This paper adopts coupon and points redemption for incentive purposes. The management of incentives including selecting the incentive type and dealing with inappropriate behavior by vehicles is beyond the scope of this paper. This is left to the states to decide as it involves legal, social, and accounting factors.

REFERENCES

[1] S. Yousefi, M. S. Mousavi, and M. Fathy, "Vehicular Ad Hoc Networks (VANETs): Challenges and Perspectives," *in Proc. the 6th International Conference on ITS Telecommunications*, Chengdu, pp. 761-766, 2006.

[2] F. Li and Y. Wang, "Routing in Vehicular Ad Hoc Networks: A Survey," *IEEE Vehicular Technology Magazine*, vol. 2, no. 2, pp. 12-22, 2007.

[3] Y. Liu, J. Bi, and J. Yang, "Research on Vehicular Ad Hoc Networks," in Proc. the 21st Annual International Conference on Chinese Control and Decision (CCDC'09), Guilin, pp. 4466-4471, 2009.

[4] H. Kabir, "Research Issues on Vehicular Ad hoc Network," *International Journal of Engineering Trends and Technology (IJETT)*, vol. 6, no. 4, pp. 174-179, 2013.

[5] S. Gillani, F. Shahzad, A. Qayyum, and R. Mehmood, "A Survey on Security in Vehicular Ad Hoc Networks," *Communication Technologies for Vehicles, Lecture Notes in Computer Science*, vol. 7865, pp 59-74, 2013.

[6] G. Samara, W. Al-Salihy, and R. Sures, "Security Analysis of Vehicular Ad Hoc Nerworks (VANET)," *in Proc. the Second International Conference on Network Applications Protocols and Services (NETAPPS)*, Kedah, pp. 55-60, 2010.

[7] M. S. Al-Kahtani, "Survey on Security Attacks in Vehicular Ad hoc Networks (VANETs)," *in Proc. the 6th International Conference on Signal Processing and Communication Systems (ICSPCS)*, Gold Coast, QLD, pp. 1-9, 2012.

[8] M. K. Jhariya, P. K. Shukla, and R. Barskhar, "Assessment of Different Attacks and Security Schemes in Vehicular Ad-hoc Network," *International Journal of Computer Applications (IJCA)*, vol. 98, no. 22, pp. 24-30, 2014.

[9] M. K. Nasir, D. Hossain, S. Hossain, M. Hasan, and B. Ali, "Security Challenges and Implementation Mechanism for Vehicular Ad Hoc Network," *International Journal of Scientific & Technology Research (IJSTR)*, vol. 2, no. 4, pp. 156-161, 2013.

 ISBN: 978-1-61208-411-4

[10] A. Agrawal, A. Garg, N. Chaudhiri, S. Gupta, D. Pandey, and T. Roy, "Security on Vehicular Ad Hoc Networks (VANET): A Review," *International Journal of Emerging Technology and Advanced Engineering (IJETAE)*, vol. 3, no. 1, pp. 231-235, 2013.

[11] C. Li, M. Hwang, and Y. Chu, "A Secure and Efficient Communication Scheme with Authenticated Key Establishment and Privacy Preserving for Vehicular Ad Hoc Networks," *Computer Communications*, vol. 31, no. 12, pp. 2803-2814, 2008.

[12] X. Lin, R. Lu, C. Zhang, H. Zhu, P. Ho, and X. Shen, "Security in Vehicular Ad Hoc Networks," *IEEE Communications Magazine*, pp. 88-95, Apr. 2008.

[13] K. Plößl and H. Federrath, "A Privacy Aware and Efficient Security Infrastructure for Vehicular Ad Hoc Networks," *Computer Standards and Interfaces*, vol. 30, pp. 390-397, 2008.

[14] M. Raya and J. Hubaux, "The security of VANETs," in Proc. the second ACM International Workshop on Vehicular Ad Hoc Networks, pp. 93-94, 2005.

[15] M. Raya and J. Hubaux, "The Security of Vehicular Ad Hoc Networks," in *Proc. the 3rd ACM Workshop on Security of Ad Hoc and Sensor Networks*, pp. 11-21, 2005.

[16] F. Sabahi, "The Security of Vehicular Ad Hoc Networks," in Proc. the 3rd International Conference on Computational Intelligence, Communication Systems, and Networks, pp. 338-342, 2011.

[17] R. Shringar, M. Kumar, and N. Singh, "Security Challenges, Issues and Their Solutions for VANET," *International Journal of Network Security & Its Applications (IJNSA)*, vol. 5, no. 5, pp. 95-105, 2013.

[18] U. M. Z. Usman and Z. B. Mohammed, "The Impact of Mobile Advertisement and Communication on Customer Relationship Management (CRM)," *in Proc. the 2012 International Conference on Economics, Business and Marketing Management*, Singapore, pp. 118-212, 2012.

[19] K. Siau and Z. Shen, "Building Customer Trust in Mobile Commerce," *Communications of the ACM*, vol. 46, no. 4, pp. 91-94, 2003.

[20] S. J. Barnes, "Wireless Digital Advertising: Nature and Implications," *International Journal of Advertising*, vol. 21, pp. 399-419, 2002.

[21] A. L. Gilbert and J. D. Kendall, "A Marketing Model for Mobile Wireless Services," *in Proc. the 36th Hawaii International Conference on System Sciences (HICSS-36)*, Hawaii, pp. 89b, 2003.

[22] S. J. Barnes and E. Scornavacca, "Mobile Marketing: The Role of Permission and Acceptance," *International Journal of Mobile Communication*, vol. 2, no. 2, pp. 128-139, 2004.

[23] P. Barwise and J. U. Farley, "The State of Interactive Marketing in Seven Countries: Interactive Marketing Comes of Age," *Journal of Interactive Marketing*, vol. 19, no. 3, pp. 67-80, 2005.

[24] S. Lee, G. Pan, J. Park, M. Gerla, and S. Lu, "Secure Incentives for Commercial Ad Dissemination in Vehicular Networks," in *Proc. the 13th Annual International Conerence on Mobile Computing and Networking*, pp. 150-159, 2007.

[25] H. Zhu, R. Lu, X. Shen, and X. Lin, "Security in Service-Oriented Vehicular Networks," *IEEE Wireless Communications*, pp. 16-22, 2009.

[26] J. Isaac, J. Camara, S. Zeadally, and J. Marquez, "A Secure Vehicle-to-Roadside Communication Payment Protocol in Vehicular Ad Hoc Networks," *Computer Communications*, vol. 31, no. 10, pp. 2478-2484, 2008.

[27] K. Daimi, M. Saed, and S. Bone, "A Multi-Level Security Architecture for Vehicular Ad Hoc Network," *in Proc. International Conference of Information Security and Internet Engineering (ICSIE'14)*, London, UK, pp. 440-455, 2014.

 ISBN: 978-1-61208-411-4

The Influences of Bridge Devices in a Scatternet Bluetooth

Celio Marcio Soares Ferreira,
Ricardo Augusto R. Oliveira,
Haroldo Santos Gambini
Computer Science Department (DECOM)
Federal University of Ouro Preto (UFOP)
Ouro Preto, Minas Gerais, Brasil
e-mail: celio@linuxplace.com.br,
{rrabelo, haroldo.santos}@gmail.com

Alejandro C. Frery
Instituto de Computação
Universidade Federal de Alagoas (UFAL)
Maceió, AL, Brasil
e-mail: acfrery@gmail.com

***Abstract*—The Adaptive Frequency-Hopping (AFH) spread spectrum do a significant influence on the Bridge devices, responsible for the inter-Piconet communication. In Bluetooth, a Bridge must stop communicating within one Piconet and must change its frequency-hopping sequence for communication with another. We propose an update of a classical linear programming Bluetooth Scatternet formation model, penalising the activation of Bridges devices, by including new constraints. This new model produced a topology, coherent with a well-known Scatternet protocol. Our improved model has an ideal distribution of data flow and power consumption similar to a well-known Scatternet protocol, $O(\log n)$ time and $O(n)$ message complexity.**

***Keywords*—bluetooth; scatternet; fhss; centralised model.**

I. Introduction

The growth in sales and manufacturing of mobile devices is a fact. Much of this is due to the great popularity of smartphones and tablets, and the possibilities of direct communication. Faced with this reality, it is necessary to simulate new scenarios and propose innovative networking solutions using the low power network interfaces integrated in most marketed devices.

Developed with a focus on low cost and low power [1], Bluetooth allows to create spontaneous network applications in environments requiring little or no user interaction. The Bluetooth communication technique is the Adaptive Frequency-Hopping (AFH) spread spectrum, a Frequency-Hopping Spread Spectrum (FHSS) variation. The AFH causes some relevant side effects during the formation of wide networks, due the Bluetooth technology characteristics and constraints [2].

The Piconet was projected for short connections and communication with low power consumption. The Piconet does not communicate using a fixed channel; all its participant nodes have the same frequency-hopping sequence coordinated by a master node and should assume master or slave roles. There is only one active master in a maximum of seven slaves connected and all of the slave data stream passes and is controlled by it. The range of a Piconet is limited by the radio power of the master node; to expand its boundaries, we have the Scatternet. It consists of a set of Piconets interconnected by a Bridge node, transmitting messages between the master nodes [3].

The Bridges nodes are the bottleneck and had higher energy consumption nodes in a Scatternet. They are responsible for all inter-Piconet communication and constantly execute a frequency-hopping synchronisation with the master nodes, making possible the devices to exchange messages in a multi-hop ad hoc wide network scenario.

The distributed algorithms for Scatternets must deal with new challenges, such as, energy limits of the devices, the different roles assumed by the nodes of the Piconet (slave, master), the Bridge nodes, Piconet traffic centralised in master node, and Bluetooth bandwidth limits. Therefore, new routing algorithms proposals and strategies of control and coordination have to be inserted to get an efficient and implementable Scatternet algorithm into a real world. As shown by Miklos et al. [4] and Jedda et al. [5], the configuration of a Scatternet has impact on the performance of the network.

To the Scatternet, there are proposals of dynamic Bluetooth Scatternet Formation (BSF) and centralised models. BSF are protocols; centralised models are a optimization models that describe a Scatternet using linear programming.

As contribuition in this work, we proposed improvements to Marsan et al. [6] centralised Scatternet model. Penalising the activation of Bridge nodes, include new constraints. This new model produced Scatternet topologies more coherent with the ones predicted by Law et al. [7], a well-known $O(\log n)$ time complexity and $O(n)$ message complexity BSF dynamic algorithm.

Section II shows the related work. In Section III, we introduce the Scatternet and its models; Section IV presents the Bridge node and its influence in efficiency of a Scatternet. In Section VI, our contribution on improving the centralised model of Marsan et al. [6] is detailed. Finally, we detail the conclusions in Section VII.

II. Related Work

The centralised model of Marsan et al. [6] provides a description of the Scatternet using linear programming. Its set of constraints are proposed in a min-max formulation resulting in a optimisation problem, solved by a centralised way. The objective is to obtain a optimal Scatternet topology that fulfils the traffic requirements and Bluetooth technology constraints, minimises the traffic load and energy consumption of the master and Bridge nodes. Therefore, this model does consider the side effects of Bluetooth frequency-hopping communicating, such as: excessive delay discovery of new nodes phase [8] and the frequently frequency resynchronisation efforts of Bridge nodes, necessary for inter-Piconets message transport.

In Law et al. [7], a new dynamic algorithm of Scatternet formation is introduced. The protocol is presented in a two-layer approach:

 ISBN: 978-1-61208-411-4

1) How the devices are organised into Scatternet;
2) How the devices can discovery each other with efficiency.

The devices are organised by sets of interconnected devices, called components, and these can be a simple device, a Piconet or a Scatternet. Each component has a leader and executes the MAIN procedure in the beginning of each round. In MAIN, the leader calls SEEK procedure with probability $(\frac{1}{3} < p < \frac{2}{3})$ and SCAN procedure. This ensures that in each component, there is at least one device performing these functions. When a leader performs SEEK it tries to acquire new slaves performing SCAN. When a device in SEEK finds a device in SCAN, the CONNECTED procedure is called, and a new link is established with the component. The reorganisation of Piconets happens by one of three operations: MOVE, MERGE, MIGRATE, and these operations ensure that each new and larger components have only one leader to coordinate the distribution of devices. The Scatternet formed by this protocol is proved $O(\log n)$ time complexity and $O(n)$ message complexity, and has the following properties:

- Any device is a member of at most two Piconets;
- The number of Piconets should be optimal, and the number of Piconets lower bound is $\lceil (n-1)/k \rceil$, being that n the number of network nodes and k the number of slaves in a Piconet.

Jedda et al. [8] analysed the impacts of changing Bluetooth parameters on the static and dynamic Scatternet formation protocols. These parameters are related to the use of the frequency hop communication technique. The Scatternet formation on static protocols happens as follows; each node alternates randomly between the **INQUIRY** and **INQUIRY SCAN** Bluetooth discovery states, when one device discovers each other, a temporally Piconet is formed until being destroyed at the end of the communication. They called this mechanism of *ALTERNATE*; see BlueStars [9]; BlueMIS [10] and BlueNet [11]. In dynamic Scatternet protocols, the discovery phase is interlaced with the network formation; the node shares its time between discovering new devices and communication in the Scatternet. The examples of dynamic protocols are: Law et al. [7] and Cuomo aet al. [12]. Jedda et al. [8] using the ns-2 [13] simulator, found that changing parameters of Bluetooth 1.2 discovery phase produces *ALTERNATE* Scatternets 3.5 times faster.

In [14], the constraints of the centralised model of Marsan et al. [6] are complemented by new discussions.

- The fact that increasing the number of Piconets that form the Scatternet hasn't benefits to the network throughput, because Bridge nodes become the communication bottlenecks;
- A discussion and proposal of a distributed algorithms in Scatternet formation, including routines for the insertion and removal of nodes.

III. SCATTERNET

The Scatternet extends the limits of a Piconet, 7 slave nodes communicating in the range of a master node coordinator, making possible a wide network using Bluetooth devices. They are collections of Piconets formed spontaneously without need of fixed infrastructure. Its coordination is complex because there is a need to cross multiple Piconets, in search of the destination and handle multiple alternate paths and cycles, following the Piconet constraints, Soares et al. [15].

Bluetooth specification does not provide details about Scatternets, and leaves open to new protocol propositions. Distributed algorithms are needed to start a Scatternets. In turn, we have different routing strategies and initialisation. These topological characteristics directly influence the flow of data over the network and energy consumption of devices.

Some examples of the challenges in creating Scatternets models:

- The need to coordinate different roles of the devices (slaves, masters and Bridges) to form a Piconet;
- Energy limitations of mobile devices;
- The low data rates of Bluetooth;
- The excessive delay during the Piconet start-up, because the side effects of AFH during the discovery of devices [15];

In the literature, we can find studies of dynamic and centralised Scatternets models.

III-A. Dynamic Model

Scatternet dynamic models are protocols, and its distributed algorithms use the following heuristic [3]:

- Any device is a member of no more than two Piconets; the number of Piconets is close to the optimal; the lower bound of Piconets is $(n-1)/k$, n being the number of network nodes and k the number of slaves in a Piconet;
- Bridge devices should never be masters. This reduces the load Scheduler of the masters, which will then only consider the intra-Piconet communication;
- The number of Piconets is restricted. This reduces the number of potential inter-Piconet conflicts in the Bridges, but limits the potential of alternative routes;
- There should be as few Piconets as possible. This reduces the number of channels to be used and thus potential interference;
- Piconets should not be connected to more than one Bridge. This minimises the coordination effort needed for Scheduling;
- A device must participate in as few Piconets as possible. This decreases the amount of inter-Piconet Scheduling in the device.

To represent a Scatternets topology, we use graph representation. It shows all the possible connections between the devices in range, and the most common Scatternet algorithms and your topologies represented by graphs are:

Trees:

- It is represented by a connected graph without cycles;
- Uses minimal edges for connection;
- There is no alternative route search between nodes;
- It is more susceptible to broken links during loss of connection or power failure of a device;
- Have more simplified routes, as there is only one possible path between nodes;
- Have a more simplified routing;
- Reduces contention in the transmission slots in TDD, thus are less susceptible to the side effects of Frequency Hopping [3];
- A minimum of Piconets is desired, making the Bridges we participate in a maximum of two.

UDG:

 ISBN: 978-1-61208-411-4

TABLE I. SCATTERNET PROTOCOLS

Index	Protocol
1	BlueStars
2	BlueMesh
3	Scatternet via Insertion and Removal of Nodes
4	BlueRings
5	Distributed Scatternet Formation Procedure (DSFP)
6	Simple Scatternet Formation
7	Scatternet via Merge, Movement and Migration
8	Scatternet Formation based on Partial Triangulation
9	BlueRing Trees
10	Scatternet Formation via grouping
11	Scatternet Formation Maintenance Extension
12	Topology Construction Protocol for Bluetooth
13	BlueTree Auto-Routing
14	Tree Scatternet
15	BlueNet
16	Blueroot and Distributed Bluetrees
17	BlueStar Islands

- An edge is defined if their Euclidean distance is greater than one;
- The graph is formed as the nodes come close.

1-Factor:

- $(n = 2)$ is expected where n is the number vertices of Piconet;
- An edge is always a set slave master.

Ring:

- The Scatternet are called Bluerings [3];
- Each device belongs and two Piconet and has two links in total; each device is master and slave at the same time;
- Supports a maximum of 2 active links; route is simplified because the packets are simply forwarded;
- A large ring can get a big delay resynchronisation, proportional to the number of Bridges.

III-A1. Scatternet protocols

The Scatternets protocols are treated as a finite state machine by most the authors [3]. They are built as mechanisms to control the relationship between the states defined by specif cation Bluettooth: **INQUIRY, INQUIRY SCAN, PAGE, PAGE SCAN,** and these states are alternate and coordinated. Some protocols also use the information for each device, such as battery capacity, type of mobile device and capacity data flow, resulting in a variety of Scatternets topologies, each with a characteristic optimisation. Table I lists the types of Scatternet protocols.

III-B. Centralised Model

The centralised model of Scatternet, also known as the static Bluetooth Scatternet model, is not a protocol. Instead, it provides a description of the Scatternet formation using mathematical programming, and constraints are proposed in a min-max formulation, leading to an optimisation problem which is solved in a centralised way. It can find the best possible performance for a given graph, obeying the Piconet Bluetooth restrictions. The objective of this model is to minimise the traffic of nodes that are subject to greater congestion and energy consumption, such as the masters and Bridges, respecting the restrictions following the full convergence of the Scatternet. After that, it can be used to generate a Scatternet formation.

For instance, the Marsan et al.'s [6] model discusses the centralised Scatternet requirements:

- Network Connectivity: there must be at least one path between two nodes in the network;
- System Complexity: in order to reduce the complexity of the network, the number of Piconets is limited to a fixed value;
- Traffic Demand: the network must support the necessary source-destination connection;
- Roles of the Node: there must be some constraints applied to some nodes, according to the role they play: master or slave.

The constraints and requirements used in this model are:

- network structure;
  - Active nodes participants of Piconet can not be greater than 8;
  - Two devices to communicate must be in the range of the other;
  - A node can only be master in a Piconet;
- system capacity: The maximum bit rate of a Piconet will equal to 1 Mbps;

IV. BRIDGE NODES

Bridges are the elements that enable multi-hop communication across the Scatternet. They are needed for inter-Piconet communication. They alternate the pattern of frequency hopping among those masters connected. The Bluetooth mode that defines this operation is the HOLD mode. This Bluetooth state is used as a solution for the coexistence of a node in more than one Piconet. During this mode the device participates in different Piconets using a Time Division Multiplexing (TDM) technique. In the Scatternet, they are implemented in two types, namely, slave-slave and master-slave.

The Bluetooth HOLD mode is used to release a connection device active with the master. During this mode, a device already connected to Piconet, can sleep for a short time allowing the master node communicate or check for new devices, this communication is called inter-Piconet.

During transport of inter-Piconets messages, the Bridge device, common to the Piconets, goes to HOLD mode. This mode allows it to switch contact among the Piconet's masters, during this process, the Bridge node needs a frequency hopping synchronisation with Piconet that wants to communicate. For this to happen, the Bridge changes its pattern of frequency hopping and begins to hear the master polls messages waiting for the moment that it may send messages from other Piconet.

Some Scatternets protocols use Bridge nodes master-slave type. This type has some limitations and performance issues. As a Piconet can only have one active master, the Bridge node that acts as the master of a Piconet and slave of the other, need to leave temporarily Piconet that acts as the master, without coordination, to be able to forward messages to another. This process has a high cost of resynchronisation, since it occurs in one of Piconets a temporary loss of its master. For most applications, this cost of excessive synchronisation prevents the use of Bridges nodes master-slave type for some Scatternets applications.

The message transport procedure between Piconets has high resynchronisation cost. The need for hopping pattern of trade of Bridge us with their masters, adjustments speed

 ISBN: 978-1-61208-411-4

and procedures for inter-Piconet scheduling to coordinate activities itinerant devices, cause this process an overhead due to the guard time slots. This is because during the exchange of communication inter-Piconet using two time slots, which during this process are unavailable for communication. These characteristics of Bridges directly influence the performance of Scatternet, as are energy consuming bottleneck points and traffic [3].

IV-1. Inter-Piconet Schedulling

In the Scatternet, to coordinate communication of Bridges algorithms, inter-Piconet scheduling is necessary. These algorithms enable the Bridge be available for communication when the master need it. They use a common solution to solve this problem, reserve slots for communication with the Bridge. These reserved slots are called Rendezvous Points (RP). Intra-Piconet Scheduling algorithms have a common feature, namely, the requirement to choose the slots and control of RP. We can list some of the approaches of inter-Piconet Scheduling [3] and protocols that deal with the rendezvous windows.

- Maximum Distance Rendezvous Point (MDRP);
- Adaptive Scheduling using a max-min RP;
- Flexible Scatternet Scheduling (FSS)
- Adaptive Presence Point Density (APPD)
- Pseudo-Random Coordinated Scatternet Scheduling (PRCSS), uses pseudo-random sequence of RP;
- Locally Coordinated Scheduling (LCS)
- QOS, the Scheduling is seen as an optimisation problem, an analysis of capacity occurs before the spread of the routes;
- Load Adaptive Algorithm (LAA), this algorithm, determines the duration of the activities Bridges of each Piconet;
- Proposal of a new JUMP mode, this new mode already has specific rules to coordinate the communication of Piconets;
- Scheduling interference analysis;
- Scheduling by analysis of theoretic queue;

IV-2. Influence of Bridge nodes in efficiency of a Scatternet

The number and position of Bridge nodes are critical for the efficiency evaluation of the resulting Scatternet topology. They are responsible for inter-Piconet communication, and are subjected to greater communication and processing overhead than other nodes.

To act as a Bridge, a Bluetooth node goes to HOLD mode; it is necessary to inter-Piconets communication. During its mode, the Bridge node awaits polls package of the masters, for the destination of messages; it have a high energy cost, because in this procedure, the device receives a computational effort of intra-Piconet Scheduling algorithm and its strategies to handle with the RP.

An efficient Scatternet topology should have a minimum number of Bridge nodes because:

1) Few Bridges means less delay for messaging between the Piconet, and less coordination effort with the master nodes;
2) A smaller number of masters in the Scatternet results in less Piconets and Bridge nodes; consequently, less synchronisation with the master nodes and performance influences of the algorithms inter-Piconet Scheduling.

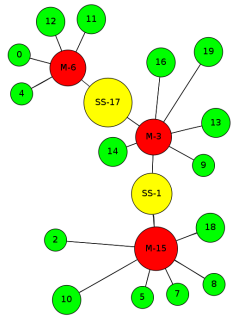


Figure 1. Common Scatternet of 20 devices found by simulation

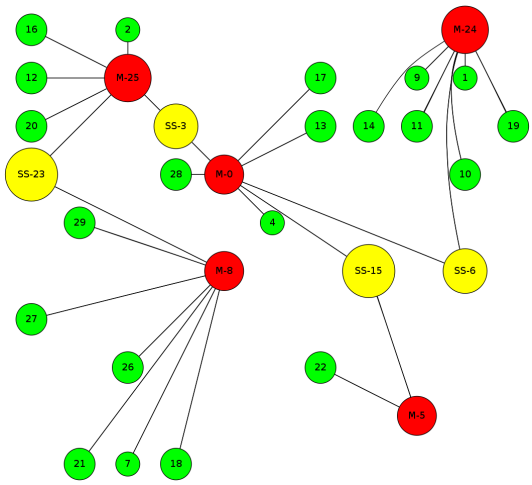


Figure 2. Common Scatternet of 30 devices found by simulation

V. TOPOLOGY ANALYSIS

To check the Scatternet topologies that would be found, we chose a dynamic and a centralized model.

V-A. Dynamic Model

To check the Scatternet topologies that would be found using a dynamic model, we chose the Law et al. [7] algorithm to simulation. This algorithm fits in item 7 of Table I, and Scatternets resulting from the simulation have $O(\log n)$ time complexity and $O(n)$ message complexity. Using ns-2 [13] with UCBT [16] extension, we simulated 30 Scatternet instances of 20 and 30 devices. With the simulation results, we generated the graph of the most common Scatternet topology found with 20 devices, in Figure 1 and with 30 devices in Figure 2.

The topology of these graphs follow the efficiency rules proposed in Section IV-2 and used by the algorithm Law et al. [7]. The red nodes are the Piconet masters M, the yellow are Bridge nodes type slave-slave (SS) and green nodes are the Piconet slaves.

V-B. Centralised Model

To find the Scatternet topologies resulting of centralized model of Marsan et al. [6], we follow the description of its

 ISBN: 978-1-61208-411-4

TABLE II. MARSAN ET AL. [6] SCATTERNET CONSTRAINTS

Constraint	Description
1	a node is either a master, or a slave or a Bridge;
2	a slave is assigned to one master at most;
3	a slave or a master are assigned to one Piconet at least; while a Bridge is assigned to two Piconets at least;
4	a master is assigned to it-self;
5	maximum connect distance is Z_{MAX};
6	limits the size of Piconet to X_{MAX};
7	If nodes i and j are masters; the assignment of i to j if is assigned to i;
8	prevents cycles among sets of three nodes;
9	the maximum number of masters is M_{MAX};
10	nodes in M to be masters;
11	nodes in set V to be slaves.

model:

The model from Marsan et al. [6] is described as follows:

- N - Number of nodes;
- C - Connections through network;
- M_{MAX} - Maximum Piconets;
- X_{MAX} - Maximum number of active nodes in Piconet;
- Z_{MAX} - Maximum radius of Piconet.
- M - Nodes constrained to act as masters;
- V - Nodes constrained to act as slaves.

For each node i, $i \in N$, three binary variables are defined: μ_i,β_i, and σ_i:

- μ_i, is equal to 1 if the node is a master and 0 otherwise;
- β_i is equal to 1 if the node is a Bridge and 0 otherwise;
- σ_i is equal to 1 if the node is a slave and 0 otherwise;

For each pair of nodes $(i, j), i, j \in N$, the set $X = \{x_{ij}\}$, x_{ij} is 1 if j is assigned to master i, otherwise 0 .

The model has the following constraints, described in Table II :

$$\mu_i + \beta_i + \sigma_i = 1, \quad \forall i \in N \tag{1}$$

$$\sum_{i \in N} x_{ij} \leq \sigma j + |N|.\beta_j + |N|.\mu_j, \quad \forall j \in N \tag{2}$$

$$\sum_{i \in N} x_{ij} \geq 2 - \sigma j - \mu_j, \quad \forall j \in N \tag{3}$$

$$x_{ii} = \mu_i, \quad \forall i \in N \tag{4}$$

$$x_{ij}.z_{ij} \leq Z_{MAX}.\mu_i, \quad \forall i, j \in N \tag{5}$$

$$\sum_{j \in N} x_{ij} \leq X_{MAX}.\mu_i, \quad \forall i \in N \tag{6}$$

$$2 + x_{ji} \geq \mu_i + \mu_j + x_{ij}, \quad \forall i, j \in N, \quad i \neq j \tag{7}$$

$$x_{ik} + x_{jk} \leq 4 - \mu_i - \mu_j - x_{ij}, \forall i, j, k \in N, i \neq j, j \neq k \tag{8}$$

$$\sum_{i \in N} \mu_i \leq M_{MAX} \tag{9}$$

$$\sum_{i \in M} \mu_i = |M| \tag{10}$$

$$\sum_{i \in V} \sigma_i = |V| \tag{11}$$

These requirements and restrictions lead to a min-max criterion that can be solve using the CPLEX solver [17].

In Marsan et al. [6] paper, to resolve a 20 devices Scatternet topology, we used the input parameters of Table III and the resulting graph is represented by Figure 3.

TABLE III. 20 DEVICES SCATTERNET - INPUT PARAMETERS

N	C	M_{MAX}	X_{MAX}	Z_{MAX}	M	$\|V\|$
20	15	4	8	$\frac{10\sqrt{2}}{3}$	$\{7, 17\}$	0

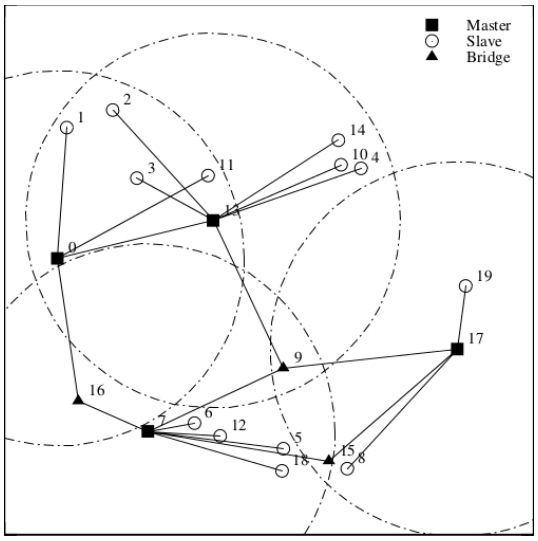


Figure 3. 20 devices Scatternet topology found in Marsan et al. [6]

TABLE IV. 30 DEVICES SCATTERNET - INPUT PARAMETERS

N	C	M_{MAX}	X_{MAX}	Z_{MAX}	M	$\|V\|$
30	4	8	8	$\frac{10\sqrt{2}}{3}$	$\{5, 25\}$	0

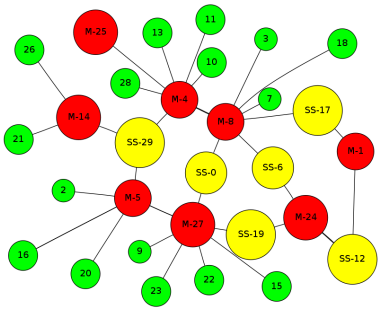


Figure 4. 30 devices Scatternet topology, found using the Marsan et al. centralised model

To resolve a 30 devices Scatternet topology with Marsan et al. centralised model, we use the input parameters of Table IV and the resulting graph is represented by Figure 4.

In the graph shown in the Figure 3, that represents the topology found how solution in the Marsan et al. [6] model, we can observe that some of the items that influence the performance of a Scatternet are neglected:

- The connection between master node 13 with node 0, is a link master / master, this setting is not possible to a Bridge node;
- Node 9 is the Bridge of three Piconets, a prohibitive result, due to the high cost of coordination with the masters conforms addressed in Section IV-1;
- We observe various network loops between the Pi-

 ISBN: 978-1-61208-411-4

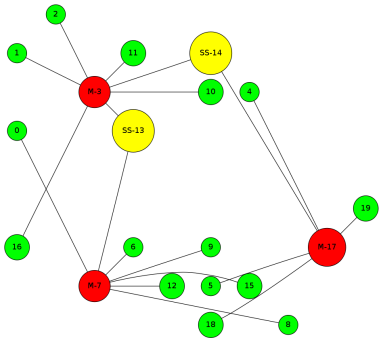


Figure 5. 20 devices Scatternet topology, found with our centralized model

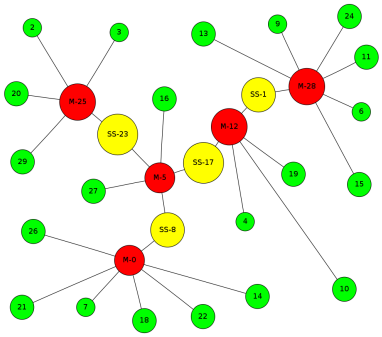


Figure 6. 30 devices Scatternet Topology, found with our centralized model.

conets of masters 7 and 17, connected by nodes 9 and 15, this increase the complexity of a Scatternet, being necessary the implementation of processing loops together with algorithms, such as the spanning tree;

- Four Piconets is an excessive amount for 20 nodes, according to Piconets lower bound proposed by Law et al. [7] and presented in Section III-A, this number should be three;

We observe that the 30 devices Scatternet topology resulted by centralised model of [6] shown in Figure 4, has the following items that influence its performance:

- The connections between nodes 4 and 25, 27 and 5, 8 and 4, are master-master links, this setting is not possible to a Piconet;
- 8 Piconets and 6 Bridges are too many for a Scatternets of 30 devices, according to Piconets lower bound proposed by Law et al. [7] and presented in Section III-A, this number should be five;

VI. Improving the Centralised Model

To get topologies similar to Scatternet protocols, the optimization models, such as Marsan et al. [6], must be improved. We added penalties to the Bridge nodes and these new constraints:

- $\mu_i + \mu_j + x_{i,j} \leq 2 \quad \forall\, i,j \in N \quad i \neq j$ **; a master must only belong to one Piconet.**
- $\beta_i + x_{ij} + x_{ji} + x_{ik} + x_{ki} + x_{kl} + x_{lk} \leq 3 \quad \forall i,j,k,l \in N \quad i \neq j \vee i \neq k \vee i \neq l \vee j < k \vee k < l$ **; a Bridge must only connect two Piconets.**

By adding penalties in Bridges and these two new constraints, we can say that the resulting graph of the solution is less prone to the effects of topology coordination delays Bridge node, responsible for inter-communication Piconet.

In order to evaluate our proposal, we use as an example the instance originally used by [6] represented by Table III and the graph of Figure 1.

We note that the Scatternets topology found with the solution of these parameters for our model, represented by graph in Figure 5, follows the heuristic of a Scatternet dynamic protocol discussed in III-A, and respects all the items needed for a efficient Scatternet discussed in Section IV-2.

In our solution, 3 masters in 3 Piconets and 2 Bridges were found, which is the same topology found by simulation of a Scatternets of 20 devices using the protocol [7] represented by Figure 1.

TABLE V. 20 DEVICES SCATTERNET - TOPOLOGY

Model	Piconets	Bridges	Piconets over the bound
Marsan et al.	4	3	1
Law et al.	3	2	0
Soares et al.	3	2	0

TABLE VI. 30 DEVICES SCATTERNET - TOPOLOGY

Model	Piconets	Bridges	Piconets over the bound
Marsan et al.	8	6	3
Law et al.	5	4	0
Soares et al.	5	4	0

These results are significant because the algorithm [7] has a cost $O(\log n)$ time complexity and $O(n)$ message complexity. Given this result, we can say that the resulting graph of our model is one Scatternet with ideal distribution data flow and energy consumption.

To validate our centralised model in larger Scatternet, we use the input parameters of Table IV. The solution of a 30 nodes Scatternet topology is represented by Figure 6.

The Scatternets found by our model has the same topology of graph Figure 2 formed by the dynamic model [7], 5 Masters in 5 Piconets and 4 Bridges. This topology follows the lower bound of Piconets proposed by [7], and has the fewest possible Bridges.

Comparing with the results of the original model from Figure 2 with the graph of our solution Figure 6, we can see fewer Piconets, 5 against 8 of the original model, fewer Bridges, 4 against 6 of the original model. Table V and Table VI summarizes our results in a comparison with the topologies found in Scatternet models, namely, centralised, dynamic and our centralised model, respectively.

VII. Conclusion and Future Work

In a scatternet, Bridges are actually points of greatest loss of efficiency because they are the nodes responsible for the coordination of inter-Piconet packet traffic. The computational effort of this process makes them network bottlenecks and points of higher power consumption by definition.

The centralised model that uses mathematical programming is useful in evaluating the performance of the simplest Scatternet topologies. In adapting the classic model of Marsan et al. [6] by changing the weights of the Bridges and adding new constraints, we achieved results similar to those obtained by

 ISBN: 978-1-61208-411-4

simulation of dynamic algorithm.

In addition, we can conclude that our resulting graph of the static Bluetooth Scatternet model represents a Scatternet with an ideal distribution of data flow and power consumption, since its result is similar to that of Law et al. [7]: complexity of $O(\log n)$ time complexity and $O(n)$ message complexity.

In ours solution, the topology found is coherent with the rules of efficiency of a Scatternet protocol, minimizing the several performance problems related to the positioning and number of Bridges.

In future works, we will propose a dynamic Bluetooth Scatternet Formation protocol that considers the impact of frequency-hopping in the Bridge nodes and inter-Piconet scheduling.

ACKNOWLEDGEMENT

Thanks to Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), to CNPq, CAPES, UFOP (Universidade Federal de Ouro Preto), SEVA Engenharia and FAPEAL.

REFERENCES

[1] Bluetooth.com, "The bluetooth network effect," http://www.bluetooth.com/Pages/network-effect.aspx/, Last Visited in 26/04/2015.

[2] bluetooth.org, "The bluetooth, bsig bluetooth specification adopted documents," https://www.bluetooth.org/en-us/specification/adopted-specifications/, Last Visited in 26/04/2015.

[3] R. M. Whitaker, L. Hodge, and I. Chlamtac, "Bluetooth scatternet formation: A survey," vol. 3, no. 4. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., Jul. 2005, pp. 403–450.

[4] G. Miklos, A. Racz, Z. Turanyi, A. Valko, and P. Johansson, "Performance aspects of bluetooth scatternet formation," in *In Proceedings of the First Annual Workshop on Mobile Ad Hoc Networking and Computing*, 2000, pp. 147–148.

[5] A. Jedda, A. Casteigts, G. Jourdan, and H. T. Mouftah, "Bluetooth scatternet formation from a time-efficiency perspective," in *Wireless Networks*, vol. 20, no. 5, 2014, pp. 1133–1156. [Online]. Available: http://dx.doi.org/10.1007/s11276-013-0664-z

[6] M. Ajmone Marsan, C. F. Chiasserini, A. Nucci, G. Carello, and L. De Giovanni, "Optimizing the topology of bluetooth wireless personal area networks," in *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2, 2002, pp. 572–579 vol. 2.

[7] C. Law, A. K. Mehta, and K.-Y. Siu, "A new bluetooth scatternet formation protocol," in *MobiHoc 2002, The ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Lausanne, Switzerland, june 2002, pp. 183–192.

[8] A. Jedda, G.-V. Jourdan, and N. Zaguia, "Some side effects of fhss on bluetooth networks distributed algorithms," in *Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications - AICCSA 2010*, ser. AICCSA '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 1–8.

[9] C. Petrioli, S. Basagni, and I. Chlamtac, "Configuring bluestars: multihop scatternet formation for bluetooth networks," in *Computers, IEEE Transactions on*, vol. 52, no. 6, June 2003, pp. 779–790.

[10] N. Zaguia, I. Stojmenovic, and Y. Daadaa, "Simplified bluetooth scatternet formation using maximal independent sets," in *Complex, Intelligent and Software Intensive Systems, 2008. CISIS 2008. International Conference on*, 2008, pp. 443–448.

[11] Z. Wang, R. J. Thomas, and Z. J. Haas, "Performance comparison of bluetooth scatternet formation protocols for multi-hop networks," vol. 15, no. 2. Hingham, MA, USA: Kluwer Academic Publishers, Feb. 2009, pp. 209–226. [Online]. Available: http://dx.doi.org/10.1007/s11276-007-0036-7

[12] F. Cuomo, T. Melodia, and I. Akyildiz, "Distributed self-healing and variable topology optimization algorithms for qos provisioning in scatternets," vol. 22, no. 7, 2004, pp. 1220–1236.

[13] NS2, "The network simulator 2 - ns2," http://www.isi.edu/nsnam/ns/, Last Visited in 26/04/2015.

[14] C. Chiasserini, M. A. Marsan, E. Baralis, and P. Garza, "Towards feasible topology formation algorithms for bluetooth-based wpans," in *36th Hawaii International Conference on System Sciences (HICSS-36 2003), CD-ROM / Abstracts Proceedings, January 6-9, 2003, Big Island, HI, USA*, 2003, p. 313. [Online]. Available: http://dx.doi.org/10.1109/HICSS.2003.1174873

[15] C. M. Soares Ferreira, R. A. Rabelo Oliveira, H. S. Gambini, A. C. Frery, S. Delabrida, and M. F. Carneiro, "A bluetooth network dynamic graph," in *AICT 2014, The Tenth Advanced International Conference on Telecommunications*, 2014, pp. 76–80.

[16] D. A. Q. Wang, "Ucbt - bluetooth extension for ns2," http://www.cs.uc.edu/~cdmc/ucbt/, University of Cincinnati, Last Visited in 26/04/2015.

[17] IBM, "Ibm ilog cplex optimizer," http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/, Last Visited in 26/04/2015.

 ISBN: 978-1-61208-411-4

2G Ultra Low Cost Mobile Phone Positioning without GPS

Cristian Anghel, Constantin Paleologu
Telecommunications Department
Politehnica University of Bucharest
Bucharest, Romania
e-mail: {canghel, pale}@comm.pub.ro

***Abstract*—This paper describes the possible location methods available in Global System for Mobile Communications (GSM) when the Ultra Low Cost (ULC) mobile phone is not equipped with a Global Positioning System (GPS) system. The proposed simplified location procedure is to be used especially in the case of emergency calls, but also in the scope of other applications.**

Keywords-positioning, emergency, low cost, GSM, EOTD

I. INTRODUCTION

In the last years, the wireless communication systems had a continuous evolution driven by the user increased requirements and expectations: more data throughput everywhere and anytime, mobility support at higher and higher speed, enhanced applications providing a huge amount of all types of information. And of course, the paid price was an increase complexity of the equipments, both on network and on user sides. So, it comes naturally to request many things from an expensive User Equipment (UE) belonging Universal Mobile Telecommunications System (UMTS) with High Speed Packet Access (HSPA+) or Long Term Evolution Advanced (LTE-A) communication systems. And one of most important aspect refers to positioning capability. Having a GPS system included in the UE allows fast positioning when the link is available and even in deep indoor situations, when the link is not available, the Assisted GPS (A-GPS) feature provides good results. So, all the commercial applications based on UE positioning work fine in almost all the cases. But one more important aspect is that the positioning may be obtained precisely and very fast during emergency calls (911 for US or 112 for Europe), and this can save human lives.

If the above mentioned characteristics correspond to latest wireless communications systems, the main question is what is happening in the case of initial GSM system? Can a 2G only ULC mobile phone, without General Packet Radio Service (GPRS) or Enhanced Data Rates for GSM Evolution (EDGE) capabilities and without a GPS system, benefit from the same positioning features as an expensive UE? To launch such a question today is not something out of interest. Although not too many 2G only mobile phones are designed anymore, there are still a lot of 2G chipsets used in Machine to Machine (M2M) applications (for example emergency systems installed on cars which activate in case of crush). Based on this remark, the goal of this paper is to list the potential positioning methods available in a 2G network and to select one of them and to present a simplified version and the obtained results when a 2G network topology is simulated.

This paper is organized as following: Section II describes the most important positioning methods available for a GSM network, Section III presents the Enhanced Observed Time Difference (EOTD) method in details and the proposed update, Section IV provides the obtained results when two cases were simulated and Section V includes the conclusions of the study.

II. POSSIBLE LOCATION METHODS USED IN GSM

The most important and also the well known location methods in GSM can be organized in two categories, based on the principles they are applying [1][2][3] and based on the place inside the network architecture where they are being executed [4].

A. Possible location methods in GSM

- Cell-ID and Timing Advance (TA) is the simplest, but also the less accurate positioning method. Cell-ID is a procedure based on knowing which cell sector the Mobile Subscriber (MS) belongs to. The sector is known only during connected state (voice or data call) and with this method no air interface resource is required to obtain cell sector information (if the user is active). Since the location is not accurate at all (a complete sector is the place where the MS may be), additional information can be added to increase the performances. Timing Advance represents the round trip delay between the MS and the serving GSM Base Transceiver Station (BTS) and it is the time MS advances its transmission with. Using Cell-ID and TA, the MS position is narrowed down to a band within a sector.
  Another approach to improve accuracy is based on an additional radio link quality indicator. The RxLEV is a GSM parameter used to describe the received signal strength. With suitable propagation models, the distance between MS and BTS can be estimated. Figure 1 depicts the stages of this Cell-ID based method.
- Time of Arrival (TOA) determines the mobile phone position based on the intersection of the distance circles. The distance is related to the propagation delay, so knowing the time on the radio link provides the distance between BTS and MS, i.e., there is a circle around BTS where the MS can be placed. 3 measurements are required, so 3 circles to be intersected (same principle applies also in GPS, but in that case circle becomes sphere), as depicted in Figure 2.

 ISBN: 978-1-61208-411-4

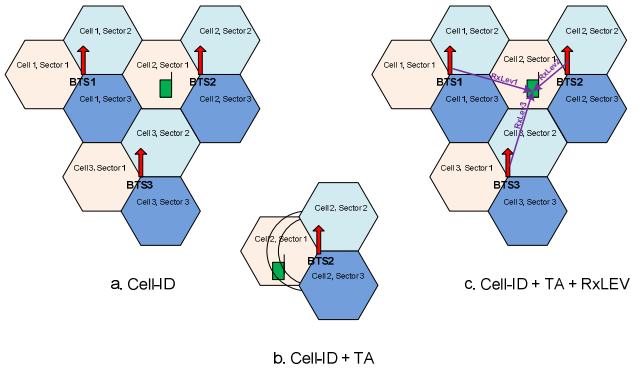


Figure 1. Cell-ID based method

- A-GPS is a method that is used when the mobile phone is equipped with a GPS system. Usually the positioning is made in this case with the information obtained from the GPS satellites. But when the Line of Sight (LOS) to the satellites does not exist, like in deep indoor situations, or when the procedure to get a fix on mobile phone position is too long, the needed information may be received via the wireless communication system.
  This method is not of interest for this paper since a 2G ULC mobile phone is considered, without having a GPS system available.
- EOTD is based on Time Difference of Arrival (TDOA), so a time difference measurement is required instead an absolute one. It is called the hyperbolic system because the time difference is converted to a constant distance difference to 2 BTSs to define a hyperbolic curve. The intersection of 2 such hyperbolas indicates the MS position, as described in Figure 3. The details of EOTD will be provided in Section III.

B. *Types of positioning methods*

- MS – based with/ without Network – assisted: for this type of methods the position is computed at MS end based on measurements performed on MS side and with/ without inputs received from network side
- Network – based with/ without MS – assisted: for this type of methods the position is computed at network end based on measurements performed on network side and with/ without inputs received from MS side

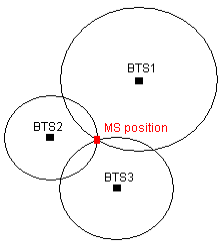


Figure 2. TOA principle

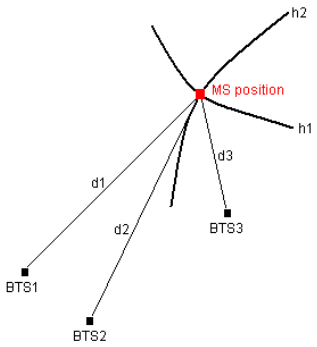


Figure 3. TDOA principle

III. ENHANCED OBSERVED TIME DIFFERENCE

A. *EOTD parameters*

There are 3 timing values required for this feature. The first one is the *Observed Time Difference* (*OTD)* and it represents the time interval observed by a MS between the bursts coming from 2 different BTSs.

The second parameter is *Real Time Difference* (*RTD)* and it represents the relative synchronization interval in the network between 2 BTSs. This time interval has to be measured by a Location Measurement Unit (LMU) on network side as the time difference between the moment when BTS_1 is sending a burst and the moment when BTS_2 is sending a burst.

The third parameter is the *Geometric Time Difference* (*GTD)* and it is the time interval measured at MS between bursts from 2 BTSs due to geometry. In other words:

$$GTD = (d_2 - d_1) / c = PD_2 - PD_1 \quad (1)$$

where d_1 and d_2 are the distances between BTS_1 and BTS_2 and the MS and c represents the speed of light. In this context PD_1 and PD_2 are the propagation delays between BTS_1 and BTS_2 and the MS.

The following relation applies:

$$GTD = OTD - RTD \quad (2)$$

In order to better understand the above explanations and the meaning of each of the 3 timing parameters, the following 2 examples are given in Figure 4. The first one considers the case when the timing relations between the 2 bursts from BTS_1 and BTS_2 is changed until MS reception due to propagation delays, while the second one keeps the same timing relation between the 2 bursts on MS reception as it was on BTSs transmission.

In Figure 4, one can observe how the timing relation between the two considered bursts was measured on LMU side and how, after the propagation delay effect was introduced by the radio channel, the bursts timing difference was observed on MS side.

 ISBN: 978-1-61208-411-4

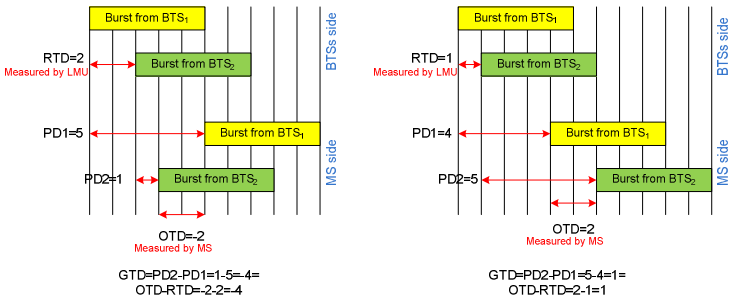


Figure 4. EOTD examples

B. How EOTD works?

In order to understand how EOTD really works, the 3GPP specifications [5] related to this feature will be also presented.

The first step of the EOTD procedure is when MS receives from network the *RTD*s values between neighbour cells and serving cell. 3GPP TS 44035, Section 4.1.1 describes the Broadcast Assistance Data received by MS from network, as presented in Table 1, with details in Section 4.1.1.12 where the Channel RTD Value IE is presented, as in Table 2.

TABLE I. EOTD ASSISTANCE DATA BROADCAST MESSAGE CONTENT

Information element	**Type/ Reference**	**Presence** (Mandatory, Conditional)
Message Structure Definition	Message Structure Definition 4.1.1.1	M
Reference Time	Reference Time 4.1.1.2	M
Ciphering Serial Number	Ciphering Serial Number 4.1.1.3	C
Time Slot Scheme	Time Slot Scheme 4.1.1.4	M
Neighbour Bitmap Definition	Neighbour Bitmap Definition 4.1.1.5	C
Sectored Channels Definition	Sectored Channels Definition 4.1.1.6	C
Sectored Channels BTS ID Definition	Sectored Channel's BTS ID Definition 4.1.1.7	C
Sectored BTS Sync/Async Definition	Sectored BTS Sync/Async Definition 4.1.1.8	C
51 Multiframe Offset Values	51 Multiframe Offset Values 4.1.1.9	M
BCC Definition	BCC Definition 4.1.1.10	M
RTD Drift Factor Values	RTD Drift Factor Values 4.1.1.11	C
Channel RTD Values	Channel RTD Values 4.1.1.12	C
Serving Cell Location	Serving Cell Location 4.1.1.13	M
Relative Neighbour Location Values	Relative Neighbour Location Values 4.1.1.14	M

TABLE II. CHANNEL RTD VALUE IE

(MSB)Varying Length (12-18 bits)(LSB)
Neighbour RTD (Last) **(MSB)**
Neighbour RTD (Last-1)
…
Neighbour RTD (2)
Neighbour RTD (1) **(LSB)**

At the second step the MS measures the *OTD* between the same neighbour cells and the serving cell from which it already received *RTD*s.

At the third step the MS computes for each neighbour cell a *GTD* related to the serving cell.

At the last step, the intersection of the obtained hyperbolas will provide the MS position (see Figure 3).

C. The math behind the TDOA principle

This section will explain in detail what computation is needed to apply the above described principle if additional BTS-MS distance information is available (extracted from TA information). The complexity of computational effort in this case is lower than the classical way of solving the problem [6]. The scope of this section is just to give an example of how this positioning problem can be solved on MS side.

Considering the notation from Figure 5, the distances between each BTS and the MS can be computed as

$$d_i = \sqrt{(x_i - x)^2 + (y_i - y)^2}\,,\ i = 1,2,3 \tag{3}$$

Subtracting the square of (3) for i=1 from the two corresponding to i=2,3 it results

$$\begin{aligned} d_j^2 - d_1^2 &= (x_j - x)^2 - (x_1 - x)^2 + \\ &+ (y_j - y)^2 - (y_1 - y)^2\ ,\ j = 2,3 \end{aligned} \tag{4}$$

If the bellow notations are used

$$\begin{aligned} &x_{j,1} = x_j - x_1 \text{ and } y_{j,1} = y_j - y_1\ ,\ j = 2,3 \\ &p_{j,1} = \frac{\left(x_j^2 + y_j^2\right) - \left(x_1^2 + y_1^2\right) + d_1^2 - d_j^2}{2}\ ,\ j = 2,3 \end{aligned} \tag{5}$$

the relations in (4) can be re-written as

$$x_{j,1}x + y_{j,1}y = p_{j,1}\ ,\ j = 2,3 \tag{6}$$

and the corresponding solution is

$$x = \frac{y_{2,1}p_{3,1} - y_{3,1}p_{2,1}}{x_{3,1}y_{2,1} - x_{2,1}y_{3,1}} \quad y = \frac{x_{3,1}p_{2,1} - x_{2,1}p_{3,1}}{x_{3,1}y_{2,1} - x_{2,1}y_{3,1}} \qquad (7)$$

IV. SIMULATION RESULTS

For simulation purposes, a 2D space between *–N*:*N* units on x axis and *–N*:*N* units on y axis was considered. *N* parameter and the value of one unit depend on the cell radius. In the below simulations *N* was considered 10 for better results analysis. In real life, the value of *N* should be aligned with the resolution provided by the network parameters (*RTD*s) and with the one of the MS measured parameters (*OTD*s). In other words, if the resolution for algorithm inputs is for example 3 meters, *N* will be chosen in such a way so that, dividing the considered distance in *N* units to obtain squares of at least 3 meters. The simulations below will show that the correct square is found. This means that the MS will be placed correctly inside a square, but the exact position in that square will remain unknown, this being the localization resolution.

In Figures 5 to 8, the BTSs are depicted with higher amplitude and with red color, while the MS is represented with lower amplitude and with green color. In both cases, Figure 5 and 7 described the considered scenario and on Figures 6 and 8 are the obtained results after applying the previously described procedure.

The first scenario corresponding to Figure 5 considered BTS_1(-1,-1), BTS_2(2,2), BTS_3(3,4) and placed the MS(7,8).

The second scenario corresponding to Figure 7 considered BTS_1(-6,8), BTS_2(-4,-7), BTS_3(8,8) and placed the MS(2,3).

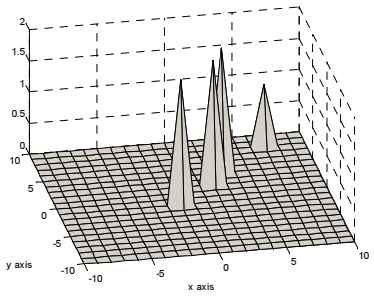


Figure 5. Simulation 1 scenario

V. CONCLUSIONS

This paper summarized the main positioning methods used in a GSM wireless communication system when no GPS module is available on MS side. It described in details the EOTD method and it provided a very simple procedure of location with low computation effort based on TA usage over the classical EOTD. Two simulated cases were presented, showing the performances of the proposed procedure. The main goal of this paper was to provide a simple positioning method suitable for 2G ULC mobile phones to be used especially in case of emergency calls, so that a tragedy effect to be limited.

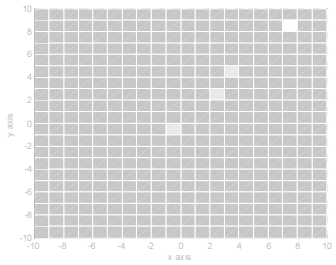


Figure 6. Simulation 1 obtained results

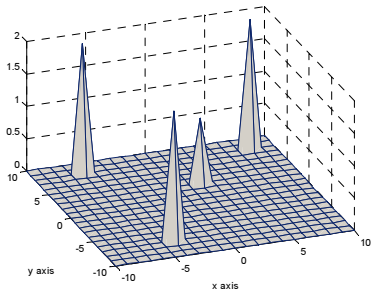


Figure 7. Simulation 2 scenario

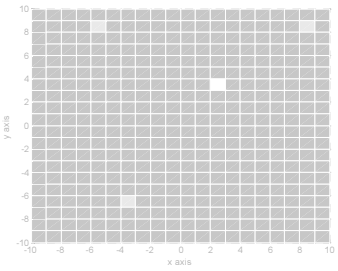


Figure 8. Simulation 2 obtained results

ACKNOWLEDGMENT

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/134398.

REFERENCES

[1] A. Melikov, "Cellular Networks – Positioning, Performance Analysis, Reliability", In Tech, 2011.

[2] A. Schmidt-Dannert, "Positioning Technologies and Mechanisms for mobile Devices", SNET2, Seminar Master Module, TU Berlin, 2010.

[3] "Overview of 2G LCS Techniques and Standards", 3GPP TSG SA2 LCS Workshop, London, January 2001.

[4] N. Deblauwe, "GSM-based Positioning: Techniques and Applications", ASP/VUBPRESS Brussels, 2008.

[5] 3GPP specifications, http://www.3gpp.org/specifications/specification-numbering, last accessed on November 2014.

[6] B. T. Fang, "Simple Solutions for Hyperbolic and Related Position Fixes", IEEE Transactions on Aerospace and Electronic Systems, vol. 26, no 5, September 1990, pp. 748–753.

 ISBN: 978-1-61208-411-4

A New Cell Selection and Handover Approach in Heterogeneous LTE Networks

Additional Criteria Based on Capacity Estimation and User Speed

Edinaldo João Costa de La-Roque, Carlos Patrick Alves da Silva, Carlos Renato Lisboa Francês
High Performance Networks Planning Laboratory – LPRAD
Federal University of Pará - UFPA
Belém, Brazil
e-mail: {elaroque, patrickalves, rfrances}@ufpa.br

***Abstract*–Long Term Evolution (LTE) heterogeneous networks represent an interesting approach to the ever increasing demand for coverage and Quality of Service (QoS) by the mobile users. Small cells play an important role in dealing with this demand by providing a means for the mobile user to overcome the problem of lack of mobile network resources or, when these resources are available, a way to dodge their poor quality in certain scenarios. However, the cell selection and handover procedures found in LTE Release 8 are inefficient in heterogeneous scenarios, since they are based only on Reference Symbol Received Power (RSRP) for cell selection and handover, and Reference Symbol Received Quality (RSRQ) for handover only parameters. In this paper, the implementation of two additional criteria is proposed as an improvement for the cell selection and handover procedures: base station capacity estimation and user speed. As the results show, the proposed algorithm has the benefit of contributing to the macrocell offloading, network load balancing, and user QoS.**

Keywords-LTE; Handover; Load Balancing; Capacity Estimation; User Speed.

I. INTRODUCTION

The explosion in the use of mobile devices and applications in recent years has led to an overload of the network infrastructure responsible for handling this traffic flow, affecting both the network performance and the user experience. To meet this growing demand for more resources, Long Term Evolution (LTE) networks are introduced as a radio access solution that provides a smooth migration path to fourth generation networks (4G), being designed to increase the capacity, coverage and speed when compared to earlier wireless systems [1]. Meanwhile, the cell selection and handover processes of the 3GPP LTE Release 8, which are based only on Reference Symbol Received Power (RSRP) and Reference Symbol Received Quality (RSRQ) [2], are inefficient because they ignore one of the main requirements demanded by the user, which is the Quality of Service (QoS) guarantee.

This paper proposes the development of additional criteria for cell selection and handover procedures in order to improve load balancing and, as a consequence, the QoS for user applications, as well. The effectiveness of these criteria is based on the ability of the base stations to send both their utilization rate estimation and base station type. Also, both User Equipment (UE) for cell selection and base station (eNodeB) for handover should be able to compute the average user speed, as this is part of the algorithms proposed in this paper.

The motivation that drives this research is the pursuit for alternative solutions for the problem of uneven load distribution over mobile networks, which leads to issues, such as call blocking and poor QoS, for example. Even though many studies have been carried out about this problem, our work brings one more contribution to the community by modeling a straightforward solution that results in macrocell offloading and network load balance by enabling low power nodes (picocells and femtocells) to take on more traffic.

In summary, our proposal has the objective of improving the overall system capacity, as well as reducing congestion by introducing a new cell selection and handover approach for LTE heterogeneous networks.

The rest of this paper is organized as follows: Section II presents some works that are related to the solution proposed in this paper. Section III introduces the basics of LTE networks with topics that are related to this paper. Section IV details the additional criteria proposed for our new cell selection and handover approach. Section V shows the main configurations adopted for the simulation environment used to validate and evaluate the proposed algorithm. In Section VI, results are discussed. Section VII presents the conclusion of the work, as well as points out to future work.

II. RELATED WORK

This section presents some works relating to the objectives of this research.

Becvar and Mach [3] presented an algorithm to mitigate the problem of redundant handovers to femtocells by estimation of throughput gain. It is stated in the paper that the gain in throughput is derived from the estimated evolution of the signals levels of all involved cells measured by the User Equipment (UE) and from an estimated time spent by the users in the Femtocell Access Point (FAP). The core of the proposed solution (estimation of throughput gain) seems to follow the idea of a kind of mapping of an RSRP

 ISBN: 978-1-61208-411-4

value to a Modulation and Coding Scheme (MCS) index value, which in turn would be translated into a maximum bit rate value that would be granted by a candidate eNodeB. The solution tries to promote user satisfaction by trying to provide throughput values as high as possible based on mitigation of redundant handovers to femtocells.

The validation of the proposal is carried out by system level simulations in MATLAB [4].

Zhang et al. [5], whose paper is referenced by [3] mentioned above, proposed a new handover algorithm based on the UE speed and QoS with the purpose of reducing unnecessary handovers.

As for the UE speed parameter, that solution classifies the speeds as:

- Low mobile state (0 to 15 km/h);
- Medium mobile state (15 to 30 km/h);
- High mobile state (above 30 km/h).

Thus, in the algorithm proposed, the handover decision process does not perform any handover to femtocells if the UE is in the high mobile state. If the UE is in the medium mobile state and the user application is not so sensitive to delay and packet loss, in contrast to applications like IPTV, VoIP, and real time games [5], then handover to femtocell is allowed. Finally, if the UE is in the low mobile state, handover to femtocell is performed.

Further, regarding the QoS parameter, that proposal basically checks if the bandwidth requirement is satisfied in order to decide if a handover will be performed or not, mainly based on cell maximum capacity and its current load.

No simulation tool was used for the validation of the proposal.

Ulvan et al. [6] presented a handover decision policy based on mobility prediction, where the position of the UEs should be known in advance. The movement prediction of the UEs is based on Markov chain transition probabilities.

Reactive and proactive handover strategies are proposed with the purpose of mitigating the frequent and unnecessary handovers in a heterogeneous mobile network scenario caused by the short coverage radius of femtocells.

Basically, the proactive handover strategy tries to estimate the characteristics of a specific position before the UE reaches that position, and that information is used so that the system can decide if it triggers a handover process or not, before a normal handover takes place. According to the authors, this strategy is expected to minimize packet loss and high latency during handover.

As for the reactive handover strategy, the handover procedure tends to be postponed as long as possible, even though a new candidate base station is discovered. The handover process is triggered only when the UE is almost losing the serving base station signal. According to the authors, this strategy is a potential mechanism to mitigate the unnecessary handovers.

The results are based on MATLAB simulation.

Compared to those works, ours promote both a relief on the network load on the system side and an improvement on the quality of service on behalf of the user by optimizing not only the handover procedure, but also the cell selection procedure, as well, by taking into account both base station capacity estimation and user speed.

Our capacity estimation method is based on the average resource availability within a period of time, as described in Section IV, while our UE speed calculation method is based on average speed in order to reduce the possibility that sudden shifts in UE speed may lead to a wrong cell selection and handover process decision making. Moreover, our solution uses a discrete event LTE simulator (OPNET Modeler), as described in Section V, which favors a more complete and realistic validation environment for the proposed algorithm.

It is worth mentioning that the use of the OPNET Modeler 17.5.A (Educational Edition) [7] forced us to adopt LTE Release 8, instead of a more up-to-date version of the mobile environment specification, which, however, does not affect the usefulness of our solution.

In short, none of the related works take into consideration average base station capacity estimation and average user speed in conjunction, so that femtocells can be dealt with accordingly, since vehicular user speed is of special importance in avoiding LTE home base stations (HeNB) from being selected or handed over to, which may contribute to service degradation for femtocell users. So, in addition to dealing properly with short radius coverage femtocells, our solution gives preference to base stations that, by estimation, have more available bandwidth resources.

These two additional criteria, adopted as proposed in Section IV, promote better load balancing by avoiding overloaded base stations from being chosen, as well as by avoiding that outdoor high speed users drain network resources from HeNB users. The result is an improvement both in the distribution of network load and user QoS, as shown in Section VI.

III. BACKGROUND

This section presents the basics of LTE networks with the topics most related to the proposal of this work.

A. LTE Heterogeneous Networks

According to Dahlman et al. [8], heterogeneous networks are a mix of cells that use different downlink transmit power, operating (partially) with the same set of carrier frequencies and with geographic coverage that overlaps, as shown in Figure 1, being also referred to as HetNets. A typical example is a picocell or a femtocell placed within the coverage area of a macrocell, as shown in Figure 1. As found in [9], in heterogeneous environment, UEs may move along the different access networks, benefitting from the different characteristics of each of them as coverage, bandwidth, latency, power consumption, costs, etc. Besides, according to 4G Americas' Board of Governor [10], in LTE networks femtocells may be office or home stations, and in the latter case they are known as Home eNodeBs (HeNB).

B. Cell Selection

As found in [11], the cell selection mechanism determines the base station that provides service to a mobile station, and this process is executed whenever the mobile

 ISBN: 978-1-61208-411-4

station joins the network (cell selection) or when the mobile device moves around in idle mode (cell reselection), as illustrated in Figure 2.

C. Handover

As it can be understood from [9], handover is an essential mechanism that guarantees mobility in a LTE network and its main function is to keep traffic flowing as the UE moves along the network. The idea behind this is simple: when a UE loses radio coverage from the serving eNodeB as it draws near another eNodeB radio coverage, a new connection has to be established to this new base station and the connection with the old one has to be undone. Therefore, handover usually happens when the serving eNodeB signal deteriorates, causing poor communication quality between the UE and the network.

Further, handover may be needed in order to promote network load balancing even if the current serving base station signal strength and quality are good. Other potential reasons to trigger a handover process is the need of the UE for better QoS, lower costs, more bandwidth, etc, which can cause the UE to search for base stations that offer better service conditions.

D. Quality of Service (QoS)

According to Sesia et al. [2], many applications may be running at the same time on the UE, each of them with its own QoS requirement. QoS is mainly about priority, packet delay, and packet loss error rate, in accordance with Table I. For instance, a UE may be on a Voice Over Internet Protocol (VoIP) phone call while navigating the Internet with a web browser and/or downloading files via File Transfer Protocol (FTP), all at the same time. While VoIP has more critical QoS requirements, such as delay and jitter, FTP file transfer requires a much lower packet loss error rate.

With the purpose of supporting multiple QoS requirements, different Evolved Packet System (EPS) bearers - logical channels that are bound to specific QoS Class Identifiers (QCIs) - are configured in the system. These EPS bearers are classified into two categories: Guaranteed Bit Rate (GBR) bearers, used for applications like VoIP, to which resources are allocated by the network in a permanent fashion, usually performed by the admission control process of an eNodeB as long as there are available system resources to establish them, and Non-Guaranteed Bit Rate (Non-GBR) bearers, which do not guarantee any particular bit rate for user applications, no permanent allocation scheme, and they are more appropriate for best effort style services like Hypertext Transfer Protocol (HTTP) and FTP.

Still according to [2], every bearer has a QCI and an Allocation and Retention Priority (ARP) bound to it. The ARP parameter is used for admission control and it decides if a certain bearer should be admitted or not, and in the case it should, it is also used to decide if a lower priority level bearer should be dropped to make room for the new one in case of network congestion.

According to Holma and Toskala [12], as part of the connection procedure of the UE to the network, an Internet Protocol (IP) address is assigned by the Packet Data Network Gateway (P-GW) to the UE and at least one bearer is established: the default bearer, which is always of Non-GBR type.

This bearer remains established for all the time period of the connection to the Packet Data Network (PDN), and it has its initial values assigned by the Mobile Management Entity (MME), a component of the Evolved Packet Core (EPC). Meanwhile, additional bearers, known as dedicated bearers, may also be established at any moment during or after the connection process is accomplished. A dedicated bearer may be a GBR or Non-GBR one.

E. Resource Allocation Mechanisms

LTE radio access makes use of a set of technologies that assures high spectral efficiency (data capacity) in its wireless interface with the UE. The main technologies adopted by LTE features high data flow in the downlink direction [12]. LTE makes use of Orthogonal Frequency-Division Multiple Access (OFDMA) in the downlink direction, whereas Single Carrier Frequency Division Multiple Access (SC-FDMA) is adopted in the uplink direction. These two technologies provide orthogonality for their subcarriers, thereby reducing interference, as well as improving network capacity. For LTE Release 8, the maximum bandwidth occupied in the frequency spectrum is 20 MHz.

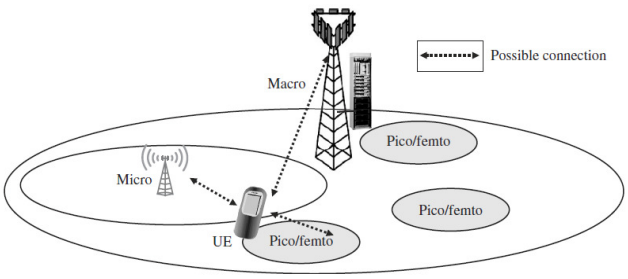


Figure 1. Example of heterogeneous network [12]

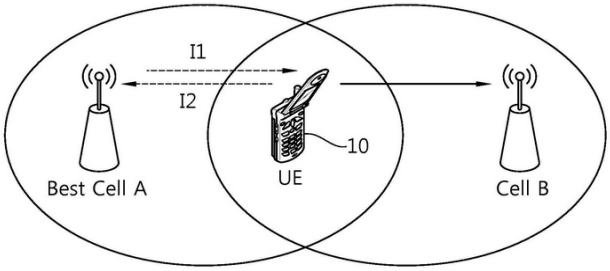


Figure 2. Cell selection and reselection [13]

 ISBN: 978-1-61208-411-4

TABLE I. LTE QCI TABLE, ADAPTED FROM [2,12]

QCI	Bearer Type	Priority	Packet Delay (ms)	Packet Loss Error Rate	Services
1	GBR	2	100	10^{-2}	Voice conversation (VoIP)
2		4	150	10^{-3}	Video conversation (live *streaming*)
3		5	300	10^{-6}	Video without conversation (buffered *streaming*)
4		3	50	10^{-3}	Real time gaming
5	Non-GBR	1	100	10^{-6}	IMS Signaling
6		7		10^{-3}	Voice, video (live *streaming*), and interactive
7		6	300	10^{-6}	Video (buffered *streaming*)
8		8			Applications based on TCP protocol like www, e-mail, chat, FTP, p2p file sharing, progressive video, etc.
9		9			

Resource allocation in frequency domain happens with a resolution of 180 KHz per Resource Block (RB) both in downlink and uplink directions. Each RB is composed of 12 subcarriers with 15 KHz bandwidth each (15 KHz subcarrier spacing). That is, 1 RB = 12 subcarriers x 15 KHz = 180 KHz, and it fits into a time slot duration of 0.5 ms, which is also equivalent to 1 Physical Resource Block (PRB). As found in [9], the resource allocation happens in every Transmit Time Interval (TTI), which corresponds to a pair of RBs (or PRBs) time interval of 1 ms. Thus, for the minimum allocated bandwidth of 1.4 MHz, 6 RBs are provided, while with the maximum allocated bandwidth of 20 MHz, 100 RBs are provided, reaching a maximum of up to 1,200 subcarriers. Table II summarizes bandwidth capacities for LTE Release 8 [14].

The data throughput that can be obtained from RBs depends on the modulation scheme, which can be Quadrature Phase Shift Keying (QPSK), 16 levels Quadrature Amplitude Modulation (16QAM) or 64 levels Quadrature Amplitude Modulation (64QAM), as well as the channel coding rate. Regarding the coding rate, as the radio condition deteriorates, the system increases the coding rate thus reducing the allocated transport block size (TBS).

Throughput also depends heavily on the number of antennas used to obtain independent transmission streams by using Multiple-Input Multiple-Output (MIMO) schemes. It is worth noting that for MIMO operations, two other parameters are used, the Rank Indicator (RI) and the Precoding Matrix Index (PMI), which will not be covered here due to lack of space.

In summary, for LTE Release 8 with a 20 MHz frequency division duplexing (FDD) bandwidth, it is possible to obtain a 150 Mbps data rate in the downlink direction when using MIMO 2x2. In the case of MIMO 4x4, LTE can provide up to 300 Mbps data rate. For the uplink direction, the peak data rate can reach up to 75 Mbps.

TABLE II. OCCUPIED BANDWIDTH, ADAPTED FROM [14]

Bandwidth (MHz)	1.4	3	5	10	15	20
No. of RBs	6	15	25	50	75	100

F. Physical Downlink Shared Channel (PDSCH)

According to [2], the PDSCH is the main data-bearing downlink channel in LTE, and it is used for all user data, as well as for the broadcasting of system information. The PDSCH channel carries data in units known as Transport Blocks (TBs), each of them corresponding to one Protocol Data Unit (PDU) from the Medium Access Control (MAC) layer. The data transmission is done during the subframe duration of 1 ms, which corresponds to 1 TTI. When the PDSCH channel is used for data transmission, one or two TBs can be transmitted per UE per subframe. For details about the PDSCH channel, please refer to [2].

IV. PROPOSED ADDITIONAL CRITERIA

This section presents the methodology adopted to perform user speed calculation and eNodeB capacity estimation, as well as it describes how cell selection and handover decision processes work, according to our proposed algorithm.

A. Overview

The purpose of the additional criteria is to promote a condition in which signal strength + quality and capacity availability, in certain proportions, may affect the cell selection and handover decision processes in such a way that preference may be given to the capacity availability parameter when choosing a serving cell, without sacrificing the connections quality. For that end, a weight of 25% for the signal strength + quality parameter against 75% for the capacity availability parameter is adopted. These weights (or proportions) were chosen from various empirical experiments and, then, they were manually assigned. Please, notice that further investigation is suggested in Section VII regarding the adoption or development of a more elaborate calculation method for the weights.

The proposed additional criteria have implied modifying the C++ source code of the UE and eNodeB models of the OPNET simulator at the LTE access stratum layer, where cell selection and handover events take place.

As highlighted in the algorithm flowchart in Figure 3.a, the cell selection process had two more decision steps added: the check for UE average speed compatibility with candidate eNodeB type and the check for the enodeB with the highest capacity availability value. Regarding the handover process, as highlighted in Figure 3.b, two more steps were added: the check for UE average speed compatibility with candidate eNodeB type and the ranking of the capacity availability estimation value, as calculated from the PDSCH channel.

B. UE Speed Calculation

A software function and a data structure were created at the UE model to calculate user speed at 1s intervals, then storing the last 10 speed samples in UE´s memory. The Euclidean distance method was used to calculate the distance traveled (in meters) for every 1 second interval, based on the (x,y) coordinates variables present in the OPNET's development environment for each UE device model (in the real world, maybe GPS coordinates would be used).

 ISBN: 978-1-61208-411-4

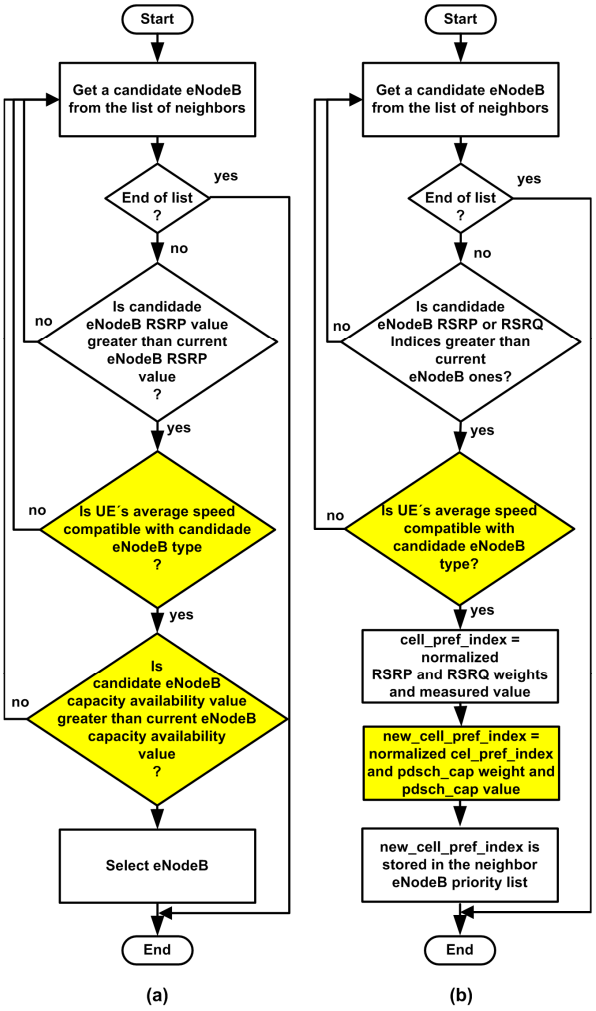


Figure 3. Flowchart of the proposed algorithm: (a) New proposed cell selection process run on UE; (b) New proposed handover process run on eNodeB

Then, the speed is obtained by dividing the calculated distance by the time spent to travel it, resulting in a speed sample expressed in m/s. Then, the arithmetic mean is calculated to obtain the average speed for the last 10 second interval, and that is the UE speed adopted by our algorithm.

Also, another software function was created to be invoked both by UE and eNodeB to calculate the average speed based on the last 10 speed samples stored on the mobile device.

C. *eNodeB Capacity Estimation via PDSCH Channel*

A software function and a data structure were created at the eNodeB model to calculate the available amount of bandwidth resources at the moment the eNodeB is assembling its radio subframes, which happens at 1 ms intervals (1 TTI). Basically, the calculation is based on the free system resources against the busy ones, as read from the PDSCH data channel. So, the last 6,000 samples (6 seconds of information) of this calculation are stored on the eNodeB´s memory and that happens for every eNodeB present on the network. So, at the moment a handover event is triggered, the serving eNodeB accesses this information from the candidate eNodeB and calculates the arithmetic mean of its available bandwidth resources for the last 6 seconds, and that information is used to decide if a handover will happen or not.

D. *New Cell Selection and Handover Decision Processes*

Besides the highest RSRP value for cell selection and RSRP (50% weight) plus RSRQ (50% weight) normalized value for handover procedures, as adopted by LTE Release 8, the proposed improvements in this paper takes into consideration UE average speed and eNodeB capacity estimation as calculated from PDSCH data traffic channel, as shown in the flowchart depicted in Figure 3.

For the handover process, according to OPNET´s source code inspection, the RSRP and RSRQ weights are normalized and applied to the normalized measurements of the RSRP and RSRQ parameters, resulting in a priority index for each neighbor eNodeB. Then, the serving eNodeB will initiate handover to the eNodeB with the highest priority index. Thus, following this idea, the capacity estimation information is also assigned a weight of 75% against the 25% weight for the eNodeB priority index (cell_pref_index in the flowchart) for handover. The purpose of assigning a 75% weight for the capacity estimation value is to make the base station resource availability value to prevail over base station signal strength and signal quality values. Then, both capacity information and its corresponding weight are normalized together with the eNodeB priority index (cell_pref_index), resulting in a new index (new_cell_pref_index) which is more influenced by the eNodeB capacity estimation value than by the signal strength and signal quality values. Then, this new index value is stored in the candidade eNodeB´s priority list for upper layer decision making relating to the handover process.

Therefore, user speed and eNodeB capacity estimation additional criteria are used in the cell selection decision process in order to avoid any UE from selecting a femtocell whenever it is in vehicular speed (above 5 km/h, for example), as it will cause another almost immediate cell reselection or handover procedure to be invoked, since the UE will soon get far from the femtocell coverage radius. These information are also used to avoid the UE from being handed over to an overloaded eNodeB whenever possible. In conjunction, both user speed and capacity estimation parameters can improve network load balancing, as well as QoS for the mobile user.

V. LTE SIMULATION ENVIRONMENT

This section presents the configuration of the LTE network scenarios used for the purposes of this work.

A. *Simulated Scenarios*

In order to validate our algorithms, three scenarios with the same LTE simulation parameters, as depicted in Figure 4, as well as detailed in Table III, were deployed on the OPNET simulator, with the following characteristics:

 ISBN: 978-1-61208-411-4

- *Baseline Scenario (REF):* Reference LTE scenario based on the standard 3GPP Release 8 specification.
- *Capacity Algorithm Scenario (CAP):* The same as the REF scenario, but with the capacity estimation algorithm enabled.
- *Capacity and Speed Algorithms Scenario (C&S):* The same as the REF scenario, but with both the capacity and the user speed algorithms enabled.

Figure 4 depicts the layout of the LTE network devices as configured in the simulator.

As for the user devices, UEs are randomly dropped, with some of them strategically placed near small cells, which is the case for femtocell users (1 stationary UE per femtocell).

The mobility profile used is random waypoint for 50 UEs with average pedestrian speed of 4.9 km/h and vehicular speed of 18 km/h.

Regarding the network traffic load, 4 stationary UEs are placed near the macrocell coverage radius in order to maintain heavy load traffic on the macrocell (4 Mbit/s high quality videoconference and 1.6 Mbit/s on-demand traffic) with the purpose of making the algorithm to give preference to less overloaded base stations (small cells) against the overloaded macrocell. Besides, 6 UEs are configured with specific trajectory, forcing the crossing of the 3 femtocells coverage radius to guarantee femtocell traffic flow, as well as to test the user speed dependency behavior of the algorithm.

Concerning the remaining user devices and their applications and QoS profile, 44 UEs run VoIP plus 3 UEs with VoIP and videoconference, all of them with bronze QoS class, 11 UEs run VoIP with silver class, and 2 UEs run VoIP and videoconference with gold class. On-demand traffic is configured as streaming multimedia of 1.6 Mbit/s in best effort mode and it is run bidirectionally between 2 UEs and the macrocell.

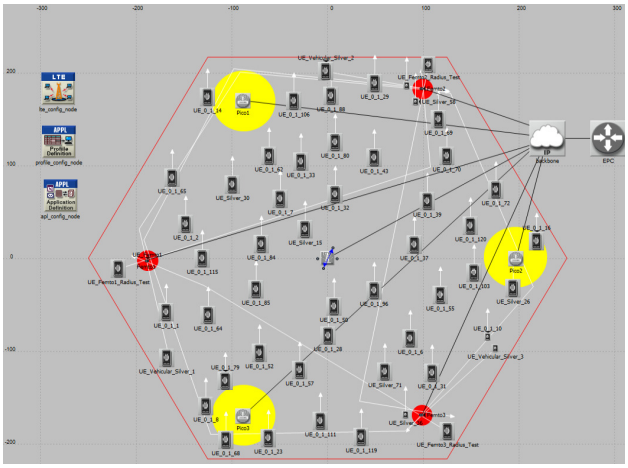


Figure 4. OPNET LTE simulated scenario layout

TABLE III. LTE SIMULATION PARAMETERS

Parameter	Macrocell	Picocell	Femtocell
No. of Base Stations	1	3	3
No. of User Devices	60 UEs Randomly Dropped		
No. of Femto Users	N/A	N/A	3
Antenna Gain	15 dBi	15 dBi	15 dBi
Max Tx Power	31 dBm	21 dBm	18 dBm
Base Station Radius	250 m	N/A	10 m
No. of Tx/Rx Antennas	2	1	1
Pathloss	Outdoor to Indoor and Pedestrian Environment (ITU-R M.1225)		UMi – Outdoor-to-indoor (ITU-R M.2135)
PHY Profile	LTE 3 MHz FDD		
Handover Type	Intra-Frequency		
Frequency Reuse	1 (2.1 GHz Carrier)		
X2 Capability	Enabled		N/A
eNodeB Selection Policy	Best Suitable eNodeB		
UE Mobility	Random Waypoint, trajectory, and fixed.		
UE Speed	4.9 km/h and 18 km/h		
User Applications	VoIP with PCM quality speech (64 Kbit/s), high quality videoconference (4 Mbit/s), and on-demand traffic (1.6 Mbit/s) .		
EPS Bearer Configuration	Bronze (QCI=6): 44 UEs with VoIP and 3 UEs with VoIP and videoconference. Silver (QCI=4): 11 UEs with VoIP, 2 of which with on-demand traffic. Gold: 2 UEs with VoIP and videoconference.		
Simulation Time	150 s with warm up time of 90 s.		

The simulation time has the duration of 150 seconds, with a warm up time of 90 seconds approximately, resulting in an effective simulation time of about 60 seconds for results collection. This simulation time is due to hardware and software constraints when simulating this realistic heavy weight LTE setup, and it was used to guarantee a stable environment at runtime, since a huge amount of events were generated during each simulation run (about 60 million events per scenario). However, after extensive work on planning and deploying variations of the given scenarios on OPNET, we concluded that the 150 s simulation time does not compromise both the algorithms behavior and the results.

B. Simulation Assumptions

In this paper, it is assumed that macro and femtocells send their respective eNodeB types, even though we implement this behavior by means of memory variables which are shared between eNodeB and UE devices. In conjunction with UE speed (vehicular or pedestrian), the eNodeB type is used to decide if a UE is allowed or not to connect to a HeNB.

Also, since OPNET Modeler 17.5.A (Educational Edition), which was used for the simulations, does not have LTE femtocell models, eNodeB models were used with femtocell parameters, instead.

 ISBN: 978-1-61208-411-4

VI. SIMULATION RESULTS

In this paper, the cell selection and handover procedures were identified as key points to be worked on in order to enhance mobile network load balance and user QoS perception, as stated in Section I. So, as an improvement proposal, besides RSRP and RSRQ parameters, two additional criteria were introduced in Section IV: capacity estimation and user speed. Then, to validate the proposed solution, the algorithms depicted in Figure 3 were developed in C++ programming language and implemented on the OPNET discrete event simulator, as well as three LTE scenarios (REF, CAP, and C&S) were planned and deployed on the simulator both for testing and statistics collection.

Thus, after extensive testing through multiple experiments performed in the realistic LTE simulation environment presented in Section V, the following metrics were chosen to evaluate the performance of the proposed algorithms:

- LTE PHY PDSCH Utilization (%);
- Total Admitted GBR Bearers;
- Total Rejected GBR Bearers;
- Downlink Dropped Packets/sec;
- LTE Delay.

The rest of this section presents the performance evaluation of the proposed solution.

A. *Network Load Balance*

From the PDSCH Utilization metric, which shows the percentage of the base stations resource utilization, as depicted in Figure 5.a, as well as numerically detailed in Tables IV and V, the load balance effect on the simulated network is clearly shown.

TABLE IV – REF SCENARIO - LTE PHY PDSCH UTILIZATION (%)

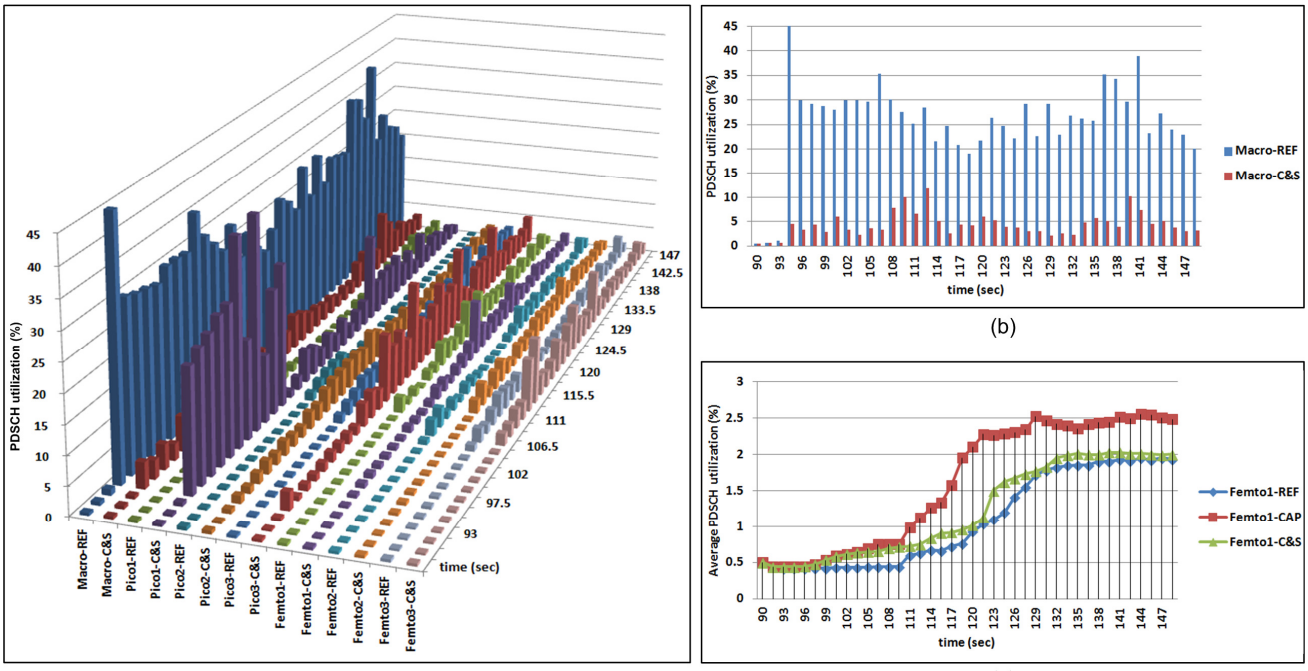
Base Station	Minimum	Average	Maximum	Std Dev
Macro	0.53603	25.465	45.059	8.6903
Pico3	0.37603	2.391	7.756	1.6539
Femto1	0.37603	1.937	6.572	1.6153
Femto3	0.35826	1.657	4.218	1.0533
Femto2	0.35826	1.380	3.969	0.9652
Pico1	0.35826	1.002	3.689	0.8392
Pico2	0.35826	0.749	2.828	0.6769

TABLE V – C&S SCENARIO - LTE PHY PDSCH UTILIZATION (%)

Base Station	Minimum	Average	Maximum	Std Dev
Pico1	0.47758	9.8548	37.542	10.136
Pico3	0.37603	5.3946	16.077	3.830
Macro	0.53603	4.4292	11.908	2.426
Femto3	0.35826	2.9768	11.703	2.719
Pico2	0.42525	2.8224	6.273	1.279
Femto1	0.37603	2.0015	9.813	1.706
Femto2	0.35826	1.8528	4.964	1.258

Figure 5.a reveals that the low power base stations (picocells and femtocells) are underutilized, while the macrocell takes on most of the network traffic in the REF scenario (standard behavior in LTE Rel-8). In contrast, in the C&S scenario (with the proposed algorithms implemented), the network traffic is offloaded from the macrocell to the low power base stations, indicating a more efficient load distribution among all the base stations. From Tables IV and V, for example, it can be seen that the macrocell had its average resource utilization decreased from 25.46 % (REF scenario) to 4.43 % (C&S scenario), which, for this specific case, represents a significant relief for the macrocell, which will have more available bandwidth for better serving UEs.

Figure 5.b highlights the macrocell traffic offloading, while the curves in Figure 5.c give an idea about the user speed influence on the cell selection and handover processes.



(a) (b) (c)

Figure 5. Load balancing effect of the proposed algorithm: (a) Network load balancing 3D visual effect; (b) Macrocell offloading effect; (c) Femto1 PDSCH utilization algorithms comparison

B. QoS Improvement

In LTE networks, the QoS guarantee for user applications is managed by the use of different EPS bearer types, as explained in Section III, item D, and it is mainly about priority, packet delay, and packet loss error rate, as shown in Table I. So, the higher the priority, as well as the lower the packet loss and the lower the network delay, the better the user QoS level.

Thus, the behavior of EPS bearers, dropped packet rate, and delay metrics is herein presented as an evidence of user QoS perception, as follows.

1) Total Admitted GBR Bearers and Total Rejected GBR Bearers

By inspecting Table VI, some conclusions can be drawn:

- There was an increase of 24.62 % (788 against 982 total bearers) in the admittance of GBR bearers, when comparing REF and C&S scenarios, which is an indication of QoS level improvement.
- There was a huge decrease in the rejection of GBR bearers (19,338 bearers from the REF scenario against 653 bearers from the C&S scenario), which is another indication of QoS level improvement.
- The huge amounts of EPS bearers (6,253, 19,338, and 26,628 bearers in the summation columns), appearing in the REF and CAP scenarios, are partially due to the excessive number of tries to establish connections to short coverage radius base stations (femtocells). An evidence of this is the lack of the user speed algorithm in the REF and CAP scenarios.

TABLE VI - GUARANTEED BIT RATE BEARERS

	Scenarios					
	Total Admitted GBR Bearers			Total Rejected GBR Bearers		
Base Station	**REF**	**CAP**	**C&S**	**REF**	**CAP**	**C&S**
Macro	434	5,355	560	495	24,975	275
Pico1	170	189	152	17,204	9	14
Pico2	44	209	135	860	350	0
Pico3	0	65	35	0	0	9
Femto1	0	14	0	0	0	0
Femto2	135	377	65	779	1,175	350
Femto3	5	44	35	0	119	5
Total	**788**	**6,253**	**982**	**19,338**	**26,628**	**653**

Figure 6 simplifies the analysis on the user QoS improvement evidence, where the best QoS case is highlighted, as shown in the graphics region corresponding to the C&S scenario, where both capacity estimation and user speed algorithms are enabled, promoting the benefit of an overall performance improvement of the system.

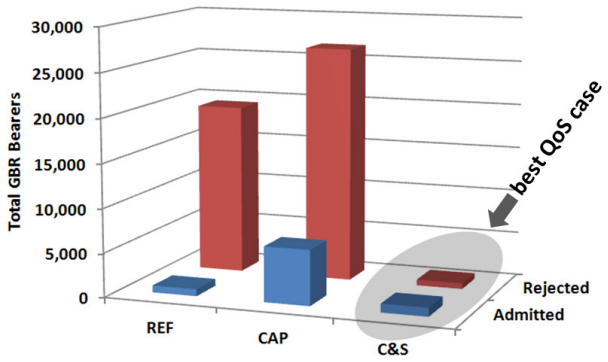


Figure 6. Best QoS case

The best QoS case is the region where 982 admitted GBR bearers meet 653 rejected GBR bearers, as highlighted in Figure 6.

2) Downlink Dropped Packets/sec

Table VII summarizes the packet loss rate for the three simulated scenarios, where it can be verified that there was a decrease of 25.35 % in the downlink packet loss rate when comparing the summations of REF and CAP scenarios, against a decrease of 24.46 % when comparing the summations of REF and C&S scenarios.

In contrast to the admitted GBR bearers versus the rejected GBR bearers analysis, conducted in B.1, where capacity estimation and user speed algorithms in conjunction were responsible for an optimum result (best QoS case), the results in Table VII indicate that the capacity estimation algorithm enabled both in CAP and C&S scenarios was indeed the responsible for the decrease in packet loss error rate, while the user speed algorithm showed a small influence of 0.89 % on this metric.

TABLE VII - DOWNLINK DROPPED PACKETS/SEC

	Scenarios		
Base Station	**REF**	**CAP**	**C&S**
Macro	2,945.90	320.68	511.55
Pico1	17.80	726.67	575.01
Pico2	12.50	544.91	475.89
Pico3	96.80	717.15	637.54
Femto1	55.00	105.97	128.78
Femto2	75.40	95.11	132.11
Femto3	64.50	96.49	164.85
Total	**3,267.90**	**2,606.98**	**2,625.73**

3) LTE Delay

Table VIII summarizes the LTE delay, which is the delay of all the traffic that flows between eNodeBs and UEs arriving at the LTE layer.

The data show that a better result was achieved with the CAP scenario (capacity estimation algorithm only), which presented an LTE delay of 1.87 seconds, against the result of 2.87 seconds for the C&S scenario (both capacity estimation

 ISBN: 978-1-61208-411-4

TABLE VIII - LTE DELAY (IN SECONDS)

	Scenarios		
Base Station	**REF**	**CAP**	**C&S**
Macro	3.54	1.20	1.83
Pico1	0.08	0.16	0.11
Pico2	0.09	0.08	0.12
Pico3	0.14	0.09	0.15
Femto1	0.38	0.11	0.38
Femto2	0.09	0.09	0.09
Femto3	0.10	0.14	0.16
Total	**4.42**	**1.87**	**2.84**

and user speed algorithms enabled). However, considering the overall system performance, as well as the other already presented metrics, the LTE delay of 2.84 seconds found in the C&S scenario still represents a significant reduction of 55.63 % (4.42 s against 2.84 s) in the LTE delay.

VII. CONCLUSION AND FUTURE WORK

This section summarizes the impact of our algorithms on the simulated LTE network, as well as give directions for future work.

The simulation results, with the adoption of the developed algorithms proposed here, showed that significant load balancing gains, as well as user QoS improvement can be achieved if the two additional criteria are adopted.

The load balancing effect of our algorithm is based on the adoption of these two additional criteria in conjunction: user speed to avoid short radius cells to be selected when user is in vehicular speed (moving too fast to benefit from a HeNB connection), as well as eNodeB capacity estimation to avoid overloaded base stations from being selected. As a consequence, QoS improvement can be achieved with our proposed solution, since the macrocell is freer to accept connections from more users. Also, femtocell users will not be impacted by users in vehicular speeds, which makes their home femtocells more available to themselves, while picocell take on more traffic load, despite of their low transmit power when compared to the macrocell.

It was demonstrated that the small cells took on more traffic flow, since the small cell users could benefit from higher modulation orders, such as 64QAM (and hence higher throughput values) for being closer to a base station with higher probability of good radio link quality. Besides, users that are closer to macrocell had more available resources at their disposal.

Femtocells had their workloads reduced mainly due to the user speed check algorithm, which caused vehicular users not to “notice” the presence of femtocells on the network. This could be seen from the reduced number of admitted GBR bearers, when the C&S scenario (capacity estimation + speed check algorithms) was compared to the REF scenario.

As future work suggestions, it is desirable to:

- Endeavour a deeper study on the weights calculation method used both for RSRP/RSRQ and capacity estimation value, so that load balancing effect can be fine tuned.
- Have a more detailed insight on the effect of outdoor UE speed on the quality of mobile service for the indoor femtocell users.
- Experiment with different path loss models and longer distances.

ACKNOWLEDGMENT

This work has been supported by National Council for the Improvement of Higher Education (CAPES), Federal University of Pará (UFPA), and Postgraduate Programme in Computer Science (PPGCC) at UFPA, Brazil.

REFERENCES

[1] C. Lin, K. Sandrasegaran, H. A. M. Ramli, and R. Basukala, "Optimized Performance Evaluation of LTE Hard Handover Algorithm with Average RSRP Constraint", International Journal of Wireless & Mobile Networks, Vol.3, No. 2, 2011, pp. 1-16.

[2] S. Sesia, I. Toufik, and M. Baker, “LTE – The UMTS Long Term Evolution: From Theory to Practice”. 2ed. United Kingdom: John Wiley & Sons, Ltd., 2011.

[3] Z. Becvar and P. Mach, “Mitigation of redundant handovers to femtocells by estimation of throughput gain”, Mobile Information Systems 9, 2013, pp. 315-330.

[4] MATLAB - The Language of Technical Computing. Available at http://www.mathworks.com/products/matlab/index-b.html. [Retrieved: May, 2015].

[5] H. Zhang, X. Wen, B. Wang, W. Zheng, and Y. Sun, “A Novel Handover Mechanism between Femtocell and Macrocell for LTE based Networks”, IEEE Second International Conference on Communication Software and Networks, 2010, pp. 228-231.

[6] A. Ulvan, R. Bestak, and M. Ulvan, "Handover Scenario and Procedure in LTE-based Femtocell Networks", UBICOMM 2010, The Fourth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, 2010, pp. 213-218.

[7] Riverbed Modeler - The fastest discrete event-simulation engine for analyzing and designing communication networks. Available at http://www.riverbed.com/products/performance-management-control/network-performance-management/network-simulation.html. [Retrieved: May, 2015].

[8] E. Dahlman, S. Parkvall, and J. Sköld, “4G LTE/LTE-Advanced for Mobile Broadband”, United States of America: Academic Press, 2011.

[9] T. Ali-Yahiya, “Understanding LTE and its Performance”, France: Springer, 2011.

[10] 4G Americas' Board of Governor, “Mobile Broadband Explosion: The 3GPP Wireless Evolution”, Electronic Publication, August 2013. Available at http://www.4gamericas.org/en/resources/white-papers/2013/. [Retrieved: May, 2015].

[11] D. Amzallag, R. Bar-Yehuda, D. Raz, and G. Scalosub, "Cell Selection in 4G Cellular Networks", Mobile Computing, IEEE Transactions on (Volume:12, Issue: 7), 2013, pp. 1443-1455.

[12] H. Holma and A. Toskala, “LTE for UMTS: Evolution to LTE-Advanced”, 2ed. Nokia Siemens Networks, Finland: John Wiley & Sons, Ltd., 2011.

[13] Apparatus and Methods of Cell Reselection in Wireless Communication System. Available at http://www.strutpatent.com/patent/08200232/apparatus-and-method-of-cell-reselection-in-wireless-communication-system#!prettyPhoto. [Retrieved: May, 2015].

[14] A. M. Ghaleb, D. Chieng, A. Ting, and A. Abdulkafi, “Throughput Performance Insights of LTE Release 8: Malaysia's Perspective”, Wireless Communication Cluster, MIMOS Berhad, Malaysia, IEEE, 2013, pp. 258-163.

 ISBN: 978-1-61208-411-4

Content Delivery Architecture for Communication Device-to-Device Wireless Networks

Charles Tim Batista Garrocho*, Mauricio Jose da Silva† and Ricardo Augusto Rabelo Oliveira‡
Department of Computer, Federal University of Ouro Preto
Ouro Preto, MG – Brazil
Email: *ctgarrocho@gmail.com, †badriciobq@gmail.com, ‡rrabelo@gmail.com

***Abstract*—The Device to Device (D2D) communications have become essential in daily life. Current technologies characteristics preclude the transparent exchange of content among devices. To address this, an architecture that manages the Wi-Fi interface device is defined. It promotes communication between devices, allowing transparent content exchange without user intervention. Two applications that employ the use of this architecture are presented. The first one, for personal devices, proved to be scalable in tests with up to nine devices. The second one, for vehicles, proved to be feasible when applied in scenarios with low speed, causing a low packet loss and high transmission rates.**

Keywords–Wireless network architecture; Management; D2D and V2I applications.

I. INTRODUCTION

The wireless communication networks have become essential in the information society. People can connect to data networks from anywhere, through various communication devices and technologies. The vehicle is a place where users spend much of their time every day [1][2].

In recent years, mobile devices such as cellphones, smartphones and tablets are gaining popularity and evolving, making the user interaction with the device a less virtual and more realistic experience. The integration of sensors, such as Bluetooth, Wi-Fi Direct, accelerometer, compass, gyroscope, microphone, camera, Global Positioning System (GPS) and radio turned simple cell phones into powerful portable machines [3][4].

A. Wireless Network Technologies

The primary means of access to the information is through cell phone networks, that allow us to have instant access to the internet services, as long as the device is located inside a cell of an antenna [5]. However, cellular networks may be lacking or fails, in case of partial or total communication infrastructure failures caused by natural disasters [6], government censorship [7], or even by interruptions in the Internet or mobile network services [8].

Although the wireless interface technologies such as Wi-Fi ad-hoc, Wi-Fi Direct and Bluetooth offer capabilities Peer-to-Peer (P2P) for information exchange in the absence of cellular network, limitations of the protocol specification, chipsets and operating systems on mobile devices make these technologies mostly useless in practice.

Current mobile devices do not support Wi-Fi ad-hoc [9], except on devices with a rooted operating system, as in [10]. Bluetooth is limited in terms of communication distance and bandwidth as well as device discovery without human interaction [11]. In addition, the Bluetooth takes a long time for pairing and most of the attempts are unsuccessful [12]. Communication via Wi-Fi Direct is another option, but the input of a Personal Identification Number (PIN) is mandatory, which demands interaction with the user, and the group formation can take up to two minutes [13].

These characteristics of the cited technologies, especially the cell phone technology, prevent the development of applications that require transparent communication, that is, the formation of the communication network and the exchange of content without the need of user interaction with the device. To solve these problems, the use and management of the Wi-Fi interface of the devices are proposed, so as to allow a transparent communication to the user.

B. Contributions

The main contribution of this paper is a content distribution architecture, where devices can become a wireless access point, or a client connected to a network provided by another device. The architecture manages the Wi-Fi interface, forms the communication network and enables data transmission transparently to the user.

As a proof of concept, the following applications have been developed:

- The first application, *Crowd Wi-Fi*, allows the exchange of information among mobile devices transparently, allowing its use in events with agglomerations of people, such as restaurants or museums;
- The second application, *Black Box*, allows the transparent exchange of information between vehicles and infrastructures installed in parking lots, which allows its use in trucks or bus fleet companies.

The results showed that the *Crowd Wi-Fi* application was able to transparently distribute content from one device to several others simultaneously. The system proved scalable, simultaneously transmitting content to 8 devices at an average transmission rate of 17Mbps.

Just like in the Wi-Fi implementation, the results of the experiments with the *Black Box* application were also encouraging. The application behaved well for data transfer between a vehicle and an infrastructure. At a distance of up to 30 meters, the system was able to deliver an average transfer rate of 500kbps, a packet loss rate of 25% and an average delay of 30ms.

The rest of the paper is organized as follows: in Section II, related works are presented. In Section III, an overview of the architecture is presented. In Section IV, the *Crowd Wi-Fi* application is presented. In Section V, the *Black Box* application is presented. In Section VI, the scenarios and metrics used for evaluation are presented. In Section VII, the results of the experiments are presented. Finally, in Section VIII, the conclusions are presented.

 ISBN: 978-1-61208-411-4

II. Related Works

The emergence and ripening of P2P content distribution significantly reduced dependence on content delivery in content distribution networks as well as bottlenecks between consumers and content providers. A lot of research regarding P2P content distribution networks has been done so far, but little has been researched on the application of P2P content distribution in wireless networks [14].

Some studies aim to optimize the latency of the response time and power consumption of the devices in wireless content distribution networks by caching the content. Boscovic et al. [15] points out that Internet access via mobile devices is increasing, and that caching content among devices can increase the availability of content and decrease the response time when accessing data.

In [16] and [17], the similarities of video content requests by mobile phone users are pointed out. The proposal is to cache the content of popular video files in smartphones and to explore the D2D communication to transmit popular videos, thus avoiding requests to the Base Station (Fixed Mobile Phone Service Station). The authors claim that their proposal improves the video transfer rate by one or two orders of magnitude.

Sharma et al. [18] developed an architecture as well as demo applications to provide communication among mobile devices in the absence or ineffectiveness of cellular infrastructure. It is presumed that at least one mobile device has cellular data connectivity, and this connectivity is shared among all devices through a mobile ad hoc network.

The work developed in [18] is the one with more similarities to the proposal of this article. Its architecture is divided into three layers: Mobile Ad Hoc Network (MANETs), Middleware Content Centric Networking (CCN) and Delay Tolerant Networking (DTN), and applications. These layers mainly enable a network abstraction to the applications. As for the proposal described in this article, it provides a description of an architecture composed of several modules. Although the architecture does not abstract the network layer, as was done in [18], applications developed with the architecture of this article allows a better use of the Wi-Fi interface functions, since it communicates directly without relying on a layer. Moreover, architecture capabilities are implemented, specified and evaluated in real network scenarios consisting of personal devices and vehicles, while most other studies only rely on simulations [15][16][17].

III. Architecture Overview

The architecture incorporates a collection of devices that, together, enable the formation and management of the content distribution network. Figure 1 illustrates the components of the architecture, which is divided into four modules. Each module has its particularity and a special function. In this architecture, a device that is a wireless access point will be called *Leader*, and the devices that connect to it will be called *Client*.

The *Main* module is the first module to be initialized in the architecture. It is responsible for running and managing the three other modules. Its first operation is to run the *Manager* module.

The *Manager* module can perform two distinct operations. The first operation consists of scanning wireless access points and, if any are found, establishing connections to them. If the connection is established, the *Slave* module is run by the *Main* module. The second operation consists in configuring the wireless access point on the device. Once the access point is configured, the *Master* module is executed by the *Main* module. The Service Set Identification (SSID) and password of the network are constant, so all devices know which wireless network they must search and connect to.

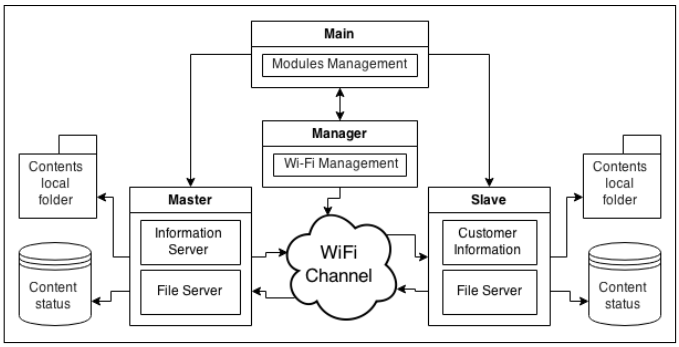


Figure 1. Schematic model of the Architecture.

The *Master* module runs exclusively on a *Leader* device, and it has information about all devices connected to it. For this reason, this module must provide information on the state of the network through an information server. On the other hand, the *Slave* module runs exclusively on a *Client* device that is connected to the *Leader*. The *Slave* module requests information about the communication network to the *Master* module running in the *Leader* device.

The contents exchanged among devices connected to the information network are transmitted through a content server on the *Master* and *Slave* modules. This server must be multitasking in order to support multiple simultaneous requests from *Clients* or *Leader* devices.

A *Leader* device simultaneously runs the *Main*, the *Manager* and the *Master* modules, while the *Client* device simultaneously runs the *Main*, the *Manager* and the *Slave* modules. In the developed applications the architecture can be used in two different ways: in the first way, one device is defined as *Leader* and all other devices as *Client*. In the second way, devices take turns acting as *Leader* and *Client*.

IV. Crowd Wi-Fi: Transparent Content Distribution Among Personal Mobile Devices

This application aims to transparently distribute content among multiple personal mobile devices in a scalable way. An example of situation where it could be used are events where there are concentrations of people, like restaurants or museums. The application was developed using the Android 4.1 operating system.

The *Crowd Wi-Fi* is divided into four modules (Table I) that follow the characteristics of the architecture. The four modules run in the background and are not affected by other applications running on the foreground on the device.

The *Main* module is the first and only Activity of the application, which takes care of the communication among modules and the management of the user interface. The *Manager* module is responsible for the device's wireless interface. The Android's Wi-Fi manager class is used both for scanning

 ISBN: 978-1-61208-411-4

TABLE I. CROWD WI-FI APPLICATION MODULES.

Module	Type	Operations
Main	Activity	Runs and manages the other modules. It's also responsible for the user interaction with the application.
Slave	AsyncTask	Requests information regarding the network to the *Master*. Provides and requests files.
Master	AsyncTask	Provides information about the state of the communication network. Provides and requests files.
Manager	Service	Manages the Wi-Fi interface and defines whether the device will be *Leader* or *Client* when communicating with the *Main*.

wireless networks and turning the device into a wireless access point.

The application's *Manager* module considers the battery level of the device for setting the duration of the the scanning for wireless networks. The lower the battery level is, the longer the scanning will be. The higher the battery level is, the shorter the scanning will be and more likely will the device become a wireless access point. The battery level is not only used to define the duration of the scanning but also when the communication network is already established. When the communication network is formed, the *Slave* module sends the device's current battery level to the the *Master* module running in the *Leader* device so it can decide which device will be the next access point should the network be destroyed (Figure 2).

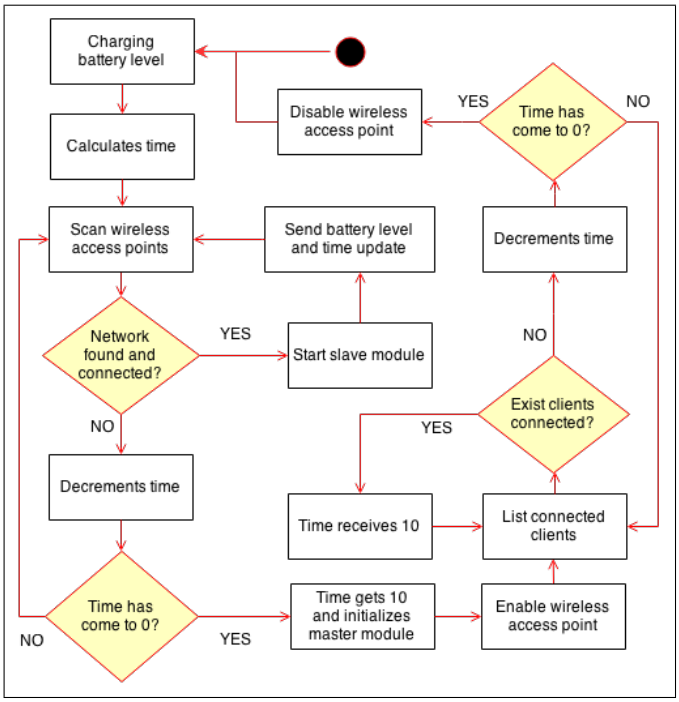


Figure 2. Activity diagram *Manager* module of the application *Crowd Wi-Fi*.

If the device becomes a *Leader*, the *Master* module is executed. This module is a AsyncTask that is triggered by the *Main* module. It is a multi threaded server that enables simultaneous connections from the *Slave* modules of the *Client* devices. The *Master* module handles three types of requests.

To make a file available on the network, the Client's *Slave* module gets the file's address and its name and builds a PUT request, sending it to the *Leader* device. The *Leader*'s *Master* module receives the Client's request and updates its local list of files available on the network as well as verifies if this new file exists on its local folder. If it does not, the *Master* module triggers a task requesting the new file to the Client's *Slave* module.

When there are no files to be made available, the *Slave* module of the *Client* devices performs two other requests to the *Leader* device, which are the LIST and the AP. In the LIST request the *Leader*'s *Master* module must respectively return lists of all the files available on the network, all the Internet Protocol (IP) addresses of *Client* devices currently connected and, finally, all device's battery levels. The *Leader*'s *Master* module sends an answer to the Client's *Slave* module containing the requested information. The Client's *Slave* module validates the list of files available on the network, and, for each file it does not have in its local folder, it triggers a task to request the file to all the devices existing in the IPs list.

After the LIST request, the *Slave* module of the *Client* device proceeds to the AP request. In this request, the *Client* device sends a message containing the device ID, its Media Access Control (MAC) address, and its current battery level. The *Leader*'s *Master* module receives this request and adds or updates this information in its local list of connected *Client* devices' battery states.

The *Master* and *Slave* modules also contain a file server. The server accepts a range of connections requesting for files and, for each request, it answers if the file exists or not. The requesting device receives the answer and if the file exists on the requested device its transmission is initialized, otherwise the connection is closed and the requesting device opens a new connection with another device, using its list of available IPs.

V. BLACK BOX: VIDEOS RECORDING MANAGEMENT AND DELIVERY IN VEHICLES

This application was developed for a truck fleet company. Each vehicle is equipped with a PandaBoardES card and a camera that monitors the driver. The main goal of this system is to record video and transmit it in chunks to a server of the truck company.

This application is divided into five modules (Table II) that follow the characteristics of the architecture presented here. The five modules run in the background, both in the vehicle and in the infrastructure.

TABLE II. BLACK BOX APPLICATION MODULES.

Module	Type	Operations
Main	ShellScript	It's the first to initialize and manages the other modules.
Camera	Python	Records videos with a specified duration and manages the available disk space.
Slave	Java	Requests for information and transmits videos to the infrastructure.
Master	Java	Offers a list of videos related to the requesting vehicle and receives videos of vehicles.
Manager	ShellScript	In the vehicle, it scans and connects to the network of the infrastructure. In the infrastructure, it configures the wireless access point for the vehicles to connect.

The *Main* module is the first to be initialized by the application and it's responsible for running and managing the five other modules. It's executed every minute by the Unix crontab tool and checks whether the other modules are running, initializing the ones that are not. To determine if a particular module is running, the application uses the Unix ps

 ISBN: 978-1-61208-411-4

tool together with the grep command with the module name to filter the results. If the operation returns nothing, the module is not running and then it is initialized.

A. Vehicle

The vehicle is only a *Client* device. It scans wireless access points and establishes connections to the ones it finds. The *Main* module of the vehicle runs and monitors the *Camera*, *Slave* and *Manager* modules.

The *Camera* module uses the camera installed in the vehicles to record videos and also manages the available disk space. It is divided into two threads. The first thread records videos from time to time in a folder. The length of each section of video is defined in a configuration file. The names of the video files are defined using the current system date and time. The second thread uses the Unix psutil tool to manage the available disk space. The maximum disk space to be used is defined in a configuration file, and if the limit is reached, the oldest video is removed from the folder. This second step is executed every minute.

The vehicle *Manager* module performs a sequence of operations, as shown in Figure 3. It uses the WPA supplicant tool to scan the wireless access points in order to verify if the vehicle is within a given cell. In order to perform the scanning, the SSID, password and Wi-Fi Protected Access (WPA) security type of the networks to be found are loaded from a configuration file. If a wireless network is found, the *Manager* module connects to it and uses the ping command to verify if the connection was established with the *Leader*. If the ping returns an error, the vehicle can not communicate with the *Leader*, in which case it runs the dhclient command on the wireless interface. The dhclient command uses the Dynamic Host Configuration Protocol (DHCP) protocol to obtain an IP address from the *Leader* and uses it to configure the wireless interface. If the ping check is successful, the vehicle checks whether the *Slave* module is running and, if not, it is initialized and its Process Identification (PID) stored, so it's not necessary to rerun it on the next interaction. The *Manager* module also monitors the *Slave* module, so it runs only when the ping test succeeds. If the ping check fails and the *Slave* module is running, it is then terminated through its PID.

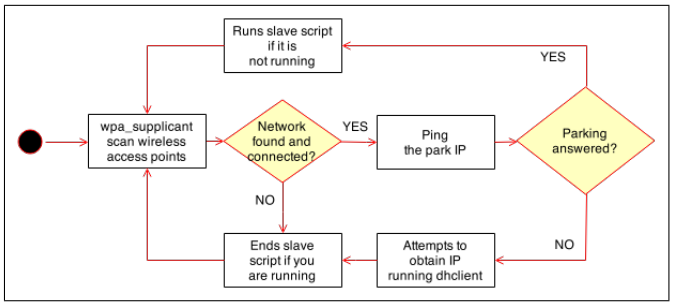


Figure 3. Activity diagram *Manager* module of the application *Black Box*.

The *Slave* module is a Java implemented client and its role is to care for the communication and management of local videos of vehicles. This module communicates with the *Master* module of the infrastructure and requests a list of videos of the vehicle. It receives the list and transmits the existing videos to the infrastructure.

B. Infrastructure

The infrastructure is nothing else than a *Leader* device, ie, a wireless access point. In the first step, an iscp-dhcp-server is set in the infrastructure's wireless card. This server manages the IP addresses of the vehicles that establish a connection to the infrastructure's wireless card. The IP address ranges and the Wi-Fi interface to be used are configured on the server.

The *Main* module of the infrastructure executes and monitors the *Master* and *Manager* modules. Unlike what occurs in the vehicle, the *Manager* module only configures the wireless access point on the infrastructure. In order to do this, it uses the hostapd tool that loads a SSID, a password and a type of WPA from a configuration file.

A sequence of message exchanges occurs between the *Master* and *Slave* modules in order to transmit the videos. The message exchange process is illustrated in Figure 4.

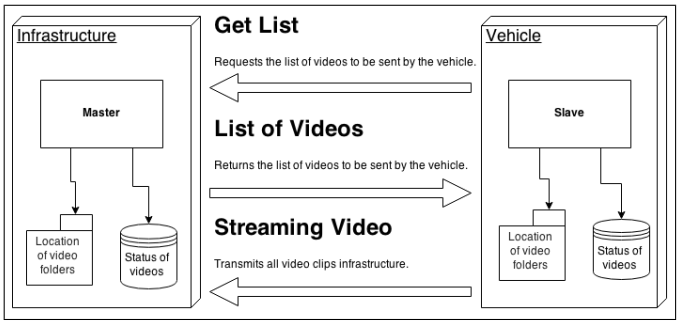


Figure 4. Communication between *Slave* and *Master* modules.

In the infrastructure, the *Master* module is a server implemented in Java and manages the videos of the vehicle. A list of videos is stored locally and contains information about the status of the videos. The infrastructure receives the list of videos of vehicles to be transmitted from the fleet company server. When a vehicle establishes connection to the Wi-Fi network of the infrastructure, the *Slave* module requests to the *Master* module a list of videos to be transmitted. The *Master* module uses a vehicle identifier to filter the videos related to it. Then, a list of videos is sent to the vehicle. The *Slave* module receives the list of videos and transmits all the video files that are requested. The *Master* module on the infrastructure receives the videos, change the videos' status and transmit them to the fleet company's server over the Internet.

VI. Evaluated Scenarios and Metrics

In this section the scenarios and metrics evaluated in the experiments are presented.

A. Crowd Wi-Fi Application

The experiments were conducted in the laboratory, in a controlled scenario in which the devices were on a table, and therefore not in movement. 9 tablet devices were used and the application was modified so that all devices requested the same 9MB file. The experiments were performed 30 times on each device. The main objective is to evaluate how the network behaves when a single device simultaneously transmits the same file to multiple devices in varying quantities.

The evaluation was performed by measuring the necessary time for the formation of a topology where the devices could communicate, the packet delay time, the packet loss rate and

 ISBN: 978-1-61208-411-4

the transmission rate. In all the experiments, the applications were executed all at once in the 9 tablets. The first experiment measured the average time spent by all the devices to establish a connection to the *Leader* and form the network topology. For the transmission delay, it was measured the time lapse between the transmission of the file and its reception on the destination. Regarding the packet loss rate, it was compared the number of packets transmitted to the number of packets actually received. Data were obtained through calculations performed on the application itself.

B. Black Box Application

The experiments were conducted on a 430 meters avenue located in the Federal University of Ouro Preto (Figure 5). A vehicle started moving from one end of the avenue (point 2), keeping the speeds of 60 km/h, 50 km/h, 40 km/h, 30 km/h and 20 km/h. The vehicle in point 1 acted as the infrastructure, standing still in the middle of the avenue. The distance between the two points was 216 meters.

Figure 5. Aerial view of the experiment region.

The evaluation of the network was performed by measuring the delay time, the loss rate and transmission rate of the packets. For the delay time it was measured the lapse between the time the packet was transmitted and the time it reached the receiver. As for the packet loss rate the number of packets transmitted was compared to the number of packets actually received. The data were obtained using the bwping software, which fired 512 bytes packets in a 2048 kbps transmission rate. Each experiment was conducted four times. The geographical positions of the vehicles were registered during the experiments.

VII. Results

In this section the results of the applications are presented.

A. Crowd Wi-Fi Application

The first experiment measured the time for devices to associate. The time was obtained through the Android application log. As stated in previous sections, 30 repetitions of this experiment were done. This experiment is important to evaluate the impact that the amount of devices has in the association time of the devices.

It can be observed in Figure 6a that when there are only a few devices, the formation time of the topology and its error rate are considerably larger. However, when the number of devices starts to increase the time to form the topology starts decreasing together with the error rate. Thus, it can be concluded that the topology formation behaves better in environments with larger numbers of devices, which makes it suitable for places with high concentrations of people, like restaurants, for instance.

The second experiment measured the packet delay time between the network communication devices. All devices stored the time when the packets were transmitted and the time when they were received in the destination. At the end, all stored times were collected and the average delay time was calculated. As stated in previous sections, 30 repetitions of this experiment were done. This experiment is important to evaluate the impact that the amount of devices has in the delay time of the packets.

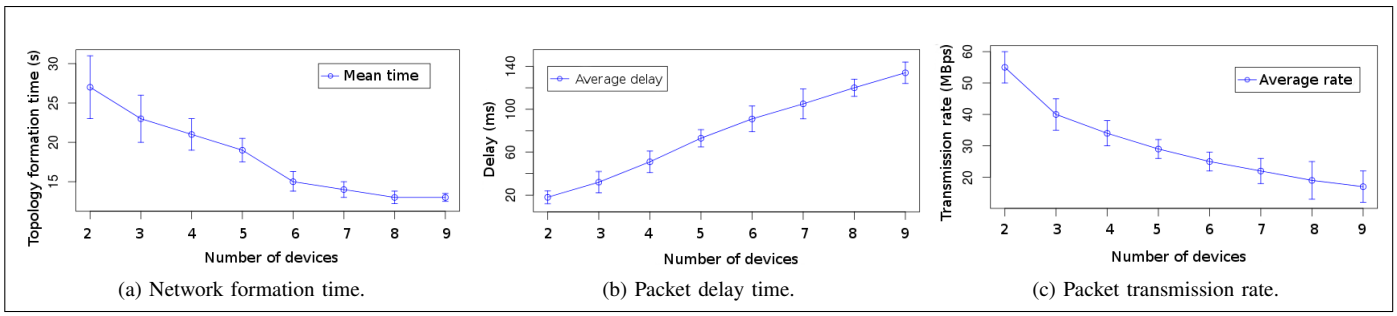
It can be observed in Figure 6b that the packet delay time increases as the number of devices that receive a file also increases. This happens because the device that transmits the file has more work to do as the communication channel is busier with more packages to be transmitted and processed at the same time.

The third experiment measured the packet transmission rate among devices on the communication network. The same process used for packet delay time was used in this experiment, but in this case, through the ages and the size of the files, it was possible to calculate the packet transmission rate.

It can be observed in Figure 6c that, like in the packet delay time, the number of devices also influences the transmission rate. The packet transmission rate decreases as the number of devices that receive the file increases. This happens because the file server on the device has more work to do as the connection bandwidth of this device is busier with multiple simultaneous connections and more packets to be processed.

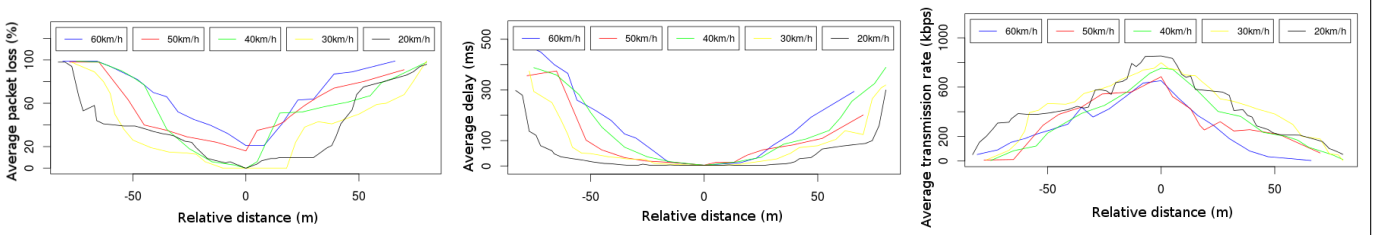
B. Black Box Application

The results were obtained from four repetitions for each experiment, and the scenario used is the one presented in Figure 5. The considered confidence interval was 95%, but it's not represented in the graph to facilitate the presentation of the information. All three experiments were evaluated at speeds of 60 km/h, 50 km/h, 40 km/h, 30 km/h and 20 km/h.



(a) Network formation time. (b) Packet delay time. (c) Packet transmission rate.

Figure 6. The results of the *Crowd Wi-Fi* application experiments.



(a) Average packet loss. (b) Average delay. (c) Average transmission rate.

Figure 7. The results of the *Black Box* application experiments.

In the graphs, the negative distance refers to the approach of the vehicle to the destination node and the positive distance refers to its distanciation.

The first experiment evaluated the packet loss rate. Figure 7a shows the loss rate. The data obtained at different speeds shows that the network behaves more robustly at lower speeds. It was possible to perform the transmission in a diameter of approximately 85 meters. The closer the vehicle is to the receiving node, the lower is the packet loss. When the nodes are at a relative distance up to 25 meters, packet loss was below 25%. The speed also impacted the packet loss, but not as much as the distance.

The second experiment evaluated the delay in the packet transmission. Figure 7b shows the delays. The delay was measured considering only the packets actually transmitted. The average delay was significantly different when considering the distance. The values obtained when the nodes were at distant points varied widely with respect to the delay obtained when the nodes were close. It was noticed that by increasing the speed, the delay in communication also suffers increase.

The third experiment evaluated the data transmission rate. Figure 7c shows the rates obtained in the communications. Data from the five experiments at different speeds showed that the average transmission rate varied over the distance.

VIII. Conclusion and Future Work

The content distribution architecture proposed in this paper successfully allowed the transparent communication in both applications developed within the Wi-Fi, showing its viability in both personal mobile devices and vehicles.

The results showed that the *Crowd Wi-Fi* application could achieve a low device association time and could also be scalable considering up to 9 devices in a communication network composed of mobile devices. Therefore, the application is feasible to be used in public places such as a restaurant, museum, or at an event where people can access content without the need for a data transmission technology.

Regarding the *Black Box* application, the results showed that below 30 km/h and at a maximum distance of 30 meters from the infrastructure, vehicles can communicate with a high transmission rate and low packet loss, making it feasible to be used for bus or truck flee companies.

In future works, besides the improvement of the applications, we intend to extend the studies, providing the *Crowd Wi-Fi* application for use in an event as well as installing the black box in a truck fleet company, with the purpose to deeply evaluate the behavior of the applications in production environments.

References

[1] P. Papadimitratos, A. L. Fortelle, K. Evenssen, R. Brignolo, and S. Cosenza, "Vehicular Communication Systems: Enabling Technologies, Applications, and Future Outlook on Intelligent Transportation", Communications Magazine IEEE, vol. 47, pp. 84–95, 2009.

[2] K. Dar, M. Bakhouya, J. Gaber, M. Wack, and P. Lorenz, "Wireless Communication Technologies for ITS Applications", Communications Magazine IEEE, vol. 48, pp. 156–162, 2010.

[3] N. D. Lane, E. Miluzzo, Hong Lu, T. Choudhury, and A. T. Campbell, "A survey of mobile phone sensing", Communications Magazine IEEE, vol. 48, pp. 140–150, 2010.

[4] R. K. Ganti, Fan Ye, and Hui Lei, "Mobile crowdsensing: current state and future challenges", Communications Magazine IEEE, vol. 49, pp. 32–39, 2011.

[5] P. Datta and S. Kaushal, "Exploration and comparison of different 4G technologies implementations: A survey", Engineering and Computational Sciences, 2014, pp. 1–6.

[6] M. Dekker and C. Karsberg, "Annual Incident Reports 2013", Technical Report October, ENISA, 2013.

[7] M. Helft and D. Barboza, "Google Shuts China Site in Dispute over Censorship", The New York Times, 22 March, 2010.

[8] T. M. Chen, "Governments and the executive 'internet kill switch"', IEEE Netw, 25 (2), 2011, pp. 2–3.

[9] IEEE Group Std, "IEEE 802.11: Wireless LAN Medium Access Control and Physical Layer Specifications", IEEE Std. 802.11, 2007.

[10] O. R. Helgason, E. A. Yavuz, S. T. Kouyoumdjieva, L. Pajevic, and G. Karlsson, "A Mobile Peer-to-Peer System for Opportunistic Content-Centric Networking", Proc. of the ACM workshop on Networking, 2010, pp. 21–26.

[11] J. C. Haartsen, "The Bluetooth radio system", IEEE Personal Communications, 2000, pp. 28–36.

[12] A. K. Pietilainen, E. Oliver, J. LeBrun, G. Varghese, and C. Diot, "MobiClique: middleware for mobile social networking", Proc. of the ACM workshop on Online social networks, 2009, pp. 49–54.

[13] Wi-Fi Alliance P2P Technical Group, "The Wi-Fi Peer-to-Peer (P2P) Technical Specification v1.0", 2009.

[14] Jin Li, "On peer-to-peer (P2P) content delivery", Peer-to-Peer Networking and Applications, 2008, pp. 45–63.

[15] D. Boscovic, F. Vakil, S. Dautovic, and M. Tosic, "Pervasive wireless CDN for greening video streaming to mobile devices", MIPRO, Proc. of the 34th International Convention, 2011, pp. 629–636.

[16] N. Golrezaei, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks", Communications (ICC), IEEE International Conference on, 2012, pp. 7077–7081.

[17] G. R. Hiertz, D. Denteneer, L. Stibor, Y. Zang, X. P. Costa, and B. Walke, "Device-to-device collaboration through distributed storage", IEEE Global Communications Conference, vol. 48, pp. 2397–2402, 2012.

[18] P. Sharma, et al., "Content and Host-Centric Information Dissemination in Delay-Tolerant Smartphone MANETs: An Architecture and Demonstration", Network Operations and Management Symposium, 2012, pp. 586–589.

 ISBN: 978-1-61208-411-4

Quasigroup Redundancy Check Codes For Safety-Critical Systems

Nataša Ilievska

Faculty of Computer Science and Engeneering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
e-mail: natasa.ilievska@finki.ukim.mk

Danilo Gligoroski

Department of Telematics
Norwegian University of Science and Technology
Trondheim, Norway
e-mail: danilog@item.ntnu.no

***Abstract*—We define error-detecting codes based on linear quasigroups. We prove that the probability of undetected errors of the defined codes, does not depend on the distribution of the characters in the input message. Next we calculate the probability of undetected errors and identify the best class of linear quasigroups of order 8 for these codes. Also, we explain how the probability of undetected errors can be controlled. At the end, we compare these codes with several CRC codes and conclude that our code has smaller probability of undetected errors than the CRC codes when code rate and block lengths are equal.**

Keywords–error-detecting codes; CRC; linear quasigroups; Safety-Critical Systems.

I. INTRODUCTION

A Cyclic Redundancy Check (CRC) is one of the most frequent mechanisms for error detection used in communication networks and storage devices. The idea presented first in 1961 in the work of Peterson and Brown [1] is for every block of data to produce a short check value attached to it. That check value is computed by an algorithm based on cyclic codes.

Very soon after their introduction, CRCs became very popular in communication and computer industry due to their mathematical simplicity and their properties to be implemented easily both in hardware and in software. Many variants of cyclic redundancy check codes have been proposed and standardized such as: CRC-8 [2], CRC-8-WCDMA (Wideband Code Division Multiple Access) [3], CRC-12 [4], CRC-ANSI (American National Standards Institute) [19], CRC-CCITT (Comité Consultatif Internationale Télégraphique et Téléphonique) [20], CRC-32 [6], CRC-64-ISO (International Organization for Standardization) [7], and many others.

Additionally, many other alternatives not based on cyclic codes have been proposed such as: Fletcher-16, Fletcher-32, Fletcher-64 [8][9] and Adler-32 [10].

Beside their typical use in digital networks, CRC codes (or their similar replacements) have been frequently used in so-called Safety-Critical Systems [11][21] that involves process control where toxic, flammable or explosive materials are used, in transportation systems such as railways, avionics and automotive systems and in nuclear power stations.

The motivation and justification of our work in this paper is closely connected with construction of redundancy check codes that will be more suitable in some use-case scenarios for those Safety-Critical Systems. This means that, while in some properties (such as the rate of the code) our codes are not that good as CRC codes, from the perspective of the probability to detect errors, our codes outperform CRCs by one or two orders of magnitude.

The paper is organized as follows. In Section II, we present the mathematical preliminaries to describe our codes. In particular it briefly defines the algebraic structures of quasigroups and linear quasigroups. In Secttion III, we describe our Linear Quasigroup Redundancy Check Codes and in Section IV, we thoroughly analyse the probability of undetected errors with our codes. In Section V, we identify a class of linear quasigroups of order 8 that give the best probabilities for error detections. In Section VI, we compare the error detection probability of our codes with three other CRC codes and we conclude the paper in Section VII.

II. MATHEMATICAL PRELIMINARIES

Previous work with error-detecting codes based on quasigroups found that the best results are obtainned with linear quasigroups [12][15].

Definition 1: Quasigroup is algebraic structure $(Q, *)$ such that

$$(\forall u, v \in Q)(\exists! x, y \in Q) \quad (x * u = v \ \& \ u * y = v) \qquad (1)$$

Definition 2: The quasigroup $(Q, *)$ of order 2^q is linear if there are non-singular binary matrices A and B of order $q \times q$ and a binary matrix C of order $1 \times q$, such that

$$(\forall x, y \in Q) \quad x * y = z \Leftrightarrow \boldsymbol{z} = \boldsymbol{x}A + \boldsymbol{y}B + C \qquad (2)$$

where $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ are binary representations of x, y and z as vectors of order $1 \times q$ and $+$ is binary addition.

When Q is a quasigroup of order 2^q, then we take that $Q = \{0, 1, ..., 2^q - 1\}$.

Example 1: One linear quasigroup of order 8 is defined with the following non-singular binary matrices

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

and the matrix $C = [0\ 0\ 0]$. In order to calculate $5 * 6$, for example, first we turn 5 and 6 into binary form. Thus, $5 = [1\ 0\ 1], 6 = [1\ 1\ 0]$ and substitute them in (2). We calculate

$$[1\ 0\ 1]A + [1\ 1\ 0]B + [0\ 0\ 0] = [1\ 1\ 1]$$

Now, [1 1 1] turned into decimal form is 7 so $5 * 6 = 7$. In the same manner we calculate all other products and obtain the following linear quasigroup:

 ISBN: 978-1-61208-411-4

$*$	0	1	2	3	4	5	6	7
0	0	3	2	1	7	4	5	6
1	7	4	5	6	0	3	2	1
2	1	2	3	0	6	5	4	7
3	6	5	4	7	1	2	3	0
4	5	6	7	4	2	1	0	3
5	2	1	0	3	5	6	7	4
6	4	7	6	5	3	0	1	2
7	3	0	1	2	4	7	6	5

Figure 1. Example of a linear quasigroup of order 8.

III. Linear Quasigroup Redundancy Check Codes

Let $(Q, *)$ be linear quasigroup of order 2^q and let $a_0a_1a_2...a_{n-1}$ be an input block of length n. The redundant characters are defined in the following way:

$$d_i = a_i * a_{i+1},\ i \in \{0, 1, ..., n-1\} \tag{3}$$

where all operations in indexes are per modulo n. This means that $d_0 = a_0 * a_1$, $d_1 = a_1 * a_2$, ..., $d_{n-2} = a_{n-2} * a_{n-1}$, $d_{n-1} = a_{n-1} * a_0$. Now, the extended message $a_0a_1a_2...a_{n-1}d_0d_1d_2...d_{n-1}$, previously turned into binary form, is transmitted through the binary symmetric channel. A block of length n is extended into a block with length $2n$, from where it follows that the rate of the code is $1/2$.

Under the influence of the noises in the channel, some of the characters may not be correctly transmitted. After receiving the block, the receiver checks if all equations (3) are satisfied. If there is some $i \in \{0, 1, ..., n-1\}$ for which the equation is not satisfied, the receiver concludes that there is an error in transmission and it asks the sender to send the block once again. But, since the redundant characters are transmitted through the binary symmetric channel, it is possible that they are incorrectly transmitted too, in a way that all equations (3) are satisfied, although some of the information characters $a_0, a_1, ..., a_{n-1}$ are incorrectly transmitted. For this reason, it is possible to have undetected errors in transmission.

IV. The Probability of Undetected Errors

With $P\{i \to j\}$ we will denote the probability that i will be transferred into j through the binary symmetric channel. The following Lemma can be easily shown.

Lemma 1: For all binary vectors $\boldsymbol{a}, \boldsymbol{b}$ i $\boldsymbol{c}$, it is true that

$$P\{\boldsymbol{a} + \boldsymbol{b} \to \boldsymbol{c} + \boldsymbol{b}\} = P\{\boldsymbol{a} \to \boldsymbol{c}\}$$

where $+$ is a binary addition on vectors.

Theorem 1: The probability of undetected errors for the considered code is independent from the distribution of the characters in the input message and form the matrix C.

Proof: Let us consider two disjoint strings of consecutive characters, say $a_ia_{i+1}...a_{i+s}$, $s \geq 0$ and $a_ja_{j+1}...a_{j+r}$, $r \geq 0$, from the input message $a_0a_1 \dots a_{n-1}$, such that there is at least one character between them, i.e., the two strings do not form a string of consecutive characters. Note that the two strings can have length one (if $s = 0$ or $r = 0$). Since, there is at least one character between the two strings, they act on different redundant characters: The string $a_ia_{i+1}...a_{i+s}$ acts on redundant characters $d_{i-1}, d_i, d_{i+1}, ..., d_{i+s}$ while the string $a_ja_{j+1}...a_{j+r}$ acts on $d_{j-1}, d_j, d_{j+1}, ..., d_{j+r}$ and this two sets of redundant characters are disjoint. For this reason, the random events:
S : the string $a_ia_{i+1}...a_{i+s}$ is incorrectly transmitted and the error is not detected;
R : the string $a_ja_{j+1}...a_{j+r}$ is incorrectly transmitted and the error is not detected;
are independent, from where it follows that $P(QR) = P(Q)P(R)$. For this reason, the probability of undetected errors will be function of the probabilities that r consecutive characters of the input message are incorrectly transmitted and the error is not detected. Therefore, in order to show that the probability of undetected errors is independent from the distribution of the characters in the input message and from the matrix C, it is enough to show that the probability that r consecutive characters of input message are incorrectly transmitted and the error is not detected is independent from the distribution of the characters in the input message and from the matrix C, for arbitrary r.

For this purpose, we introduce the following random events:
A_i : Exactly i consecutive characters from the input message $a_0a_1...a_{n-1}$ are incorrectly transmitted and the error is not detected, $i = 1, 2, ..., n$.

First, let calculate the probability $P(A_1)$, i.e., the probability that exactly one character (let say a_i) is incorrectly transmitted and the error is not detected.
Let H_j be the random event: the true value of a_i is j, $j = 0, 1, 2, ..., 2^q - 1$.

Then, using the formula for total probability, we obtain:

$$P(A_1) = \sum_{j=0}^{2^q-1} P(A_1|H_j)P(H_j) \tag{4}$$

$$\begin{aligned}
&P(A_1|H_j) = \\
&= \sum_{\substack{k=0 \\ k \neq j}}^{2^q-1} P\{a_i \to k\}P\{d_{i-1} \to a_{i-1} * k\}P\{d_i \to k * a_{i+1}\} \\
&= \sum_{\substack{k=0 \\ k \neq j}}^{2^q-1} P\{j \to k\}P\{a_{i-1} * j \to a_{i-1} * k\}\cdot \\
&\qquad \cdot P\{j * a_{i+1} \to k * a_{i+1}\} \\
&= \sum_{\substack{k=0 \\ k \neq j}}^{2^q-1} P\{\boldsymbol{j} \to \boldsymbol{k}\}P\{\boldsymbol{a_{i-1}}A + \boldsymbol{j}B + C \to \boldsymbol{a_{i-1}}A + \boldsymbol{k}B + C\}\cdot \\
&\qquad \cdot P\{\boldsymbol{j}A + \boldsymbol{a_{i+1}}B + C \to \boldsymbol{k}A + \boldsymbol{a_{i+1}}B + C\} \\
&= \sum_{\substack{k=0 \\ k \neq j}}^{2^q-1} P\{\boldsymbol{j} \to \boldsymbol{k}\}P\{\boldsymbol{j}B \to \boldsymbol{k}B\}P\{\boldsymbol{j}A \to \boldsymbol{k}A\} \\
&= \sum_{\substack{k=0 \\ k \neq j}}^{2^q-1} P\{\boldsymbol{0} \to \boldsymbol{k} + \boldsymbol{j}\}P\{\boldsymbol{0} \to (\boldsymbol{k} + \boldsymbol{j})B\}P\{0 \to (\boldsymbol{k} + \boldsymbol{j})A\}
\end{aligned} \tag{5}$$

In the last two equations in (5), Lemma 1 is used. We introduce replacement $\boldsymbol{l} = \boldsymbol{k} + \boldsymbol{j}$ in the last expression of (5). Since j is fixed and k runs through all values from $\{0, 1, ..., 2^q - 1\} \setminus \{j\}$, l will run through all values from $\{0, 1, ..., 2^q - 1\} \setminus \{0\} = \{1, ..., 2^q - 1\}$:

$$P(A_1|H_j) = \sum_{l=1}^{2^q-1} P\{\boldsymbol{0} \to \boldsymbol{l}\}P\{\boldsymbol{0} \to \boldsymbol{l}B\}P\{\boldsymbol{0} \to \boldsymbol{l}A\}, \quad \forall j \in \{0, 1, 2, ..., 2^q - 1\} \tag{6}$$

Form (4) and (6) it follows that

$$P(A_1) = \sum_{l=1}^{2^q-1} P\{\mathbf{0} \to \boldsymbol{l}\}P\{\mathbf{0} \to \boldsymbol{l}B\}P\{\mathbf{0} \to \boldsymbol{l}A\} \tag{7}$$

Equation (7) means that $P(A_1)$ is independent from the true values of a_{i-1}, a_i and a_{i+1}, i.e., it is independent from the distribution of the characters in the input message. Also, $P(A_1)$ does not depend on the matrix C.

Similarly, we derive a formula for $P(A_2)$ - the probability that exactly two consecutive characters (let say a_i and a_{i+1}) form the input message are incorrectly transmitted and the error is not detected. We introduce the random events
H_{jk} : the true value of a_i is j and the true value of a_{i+1} is k, $j, k = 0, 1, 2, ..., 2^q - 1$. Then,

$$P(A_2) = \sum_{j=0}^{2^q-1}\sum_{k=0}^{2^q-1} P(A_2|H_{jk})P(H_{jk}) \tag{8}$$

$$\begin{aligned}
&P(A_2|H_{jk}) = \\
&= \sum_{\substack{l=0\\ l\neq j}}^{2^q-1}\sum_{\substack{s=0\\ s\neq k}}^{2^q-1} P\{a_i \to l, a_{i+1} \to s\}P\{d_{i-1} \to a_{i-1} * l\}\cdot \\
&\qquad \cdot P\{d_i \to l * s\}P\{d_{i+1} \to s * a_{i+2}\} \\
&= \sum_{\substack{l=0\\ l\neq j}}^{2^q-1}\sum_{\substack{s=0\\ s\neq k}}^{2^q-1} P\{j \to l, k \to s\}P\{a_{i-1} * j \to a_{i-1} * l\}\cdot \\
&\qquad \cdot P\{j * k \to l * s\}P\{k * a_{i+2} \to s * a_{i+2}\} \\
&= \sum_{\substack{l=0\\ l\neq j}}^{2^q-1}\sum_{\substack{s=0\\ s\neq k}}^{2^q-1} P\{\boldsymbol{j} \to \boldsymbol{l}\}P\{\boldsymbol{k} \to \boldsymbol{s}\}\cdot \\
&\qquad \cdot P\{\boldsymbol{a_{i-1}}A + \boldsymbol{j}B + C \to \boldsymbol{a_{i-1}}A + \boldsymbol{l}B + C\}\cdot \\
&\qquad \cdot P\{\boldsymbol{j}A + \boldsymbol{k}B + C \to \boldsymbol{l}A + \boldsymbol{s}B + C\}\cdot \\
&\qquad \cdot P\{\boldsymbol{k}A + \boldsymbol{a_{i+2}}B + C \to \boldsymbol{s}A + \boldsymbol{a_{i+2}}B + C\} \\
&= \sum_{\substack{l=0\\ l\neq j}}^{2^q-1}\sum_{\substack{s=0\\ s\neq k}}^{2^q-1} P\{\boldsymbol{j} \to \boldsymbol{l}\}P\{\boldsymbol{k} \to \boldsymbol{s}\}P\{\boldsymbol{j}B \to \boldsymbol{l}B\}\cdot \\
&\qquad \cdot P\{\boldsymbol{j}A + \boldsymbol{k}B \to \boldsymbol{l}A + \boldsymbol{s}B\}P\{\boldsymbol{k}A \to \boldsymbol{s}A\} = \\
&= \sum_{\substack{l=0\\ l\neq j}}^{2^q-1}\sum_{\substack{s=0\\ s\neq k}}^{2^q-1} P\{\mathbf{0} \to \boldsymbol{l} + \boldsymbol{j}\}P\{\mathbf{0} \to \boldsymbol{s} + \boldsymbol{k}\}\cdot \\
&\qquad \cdot P\{\mathbf{0} \to (\boldsymbol{l} + \boldsymbol{j})B\}\cdot \\
&\qquad \cdot P\{\mathbf{0} \to (\boldsymbol{l} + \boldsymbol{j})A + (\boldsymbol{s} + \boldsymbol{k})B\}\cdot \\
&\qquad \cdot P\{\mathbf{0} \to (\boldsymbol{s} + \boldsymbol{k})A\}
\end{aligned} \tag{9}$$

We introduce replacement: $\boldsymbol{t} = \boldsymbol{l} + \boldsymbol{j}$ and $\boldsymbol{r} = \boldsymbol{s} + \boldsymbol{k}$ in the last expression of (9). When l gets all values from $Q \setminus \{j\}$, t will get all values from $Q \setminus \{0\}$. Similarly, when s gets all values from $Q \setminus \{k\}$, r will get all values from $Q \setminus \{0\}$. We obtain:

$$\begin{aligned}
&P(A_2|H_{jk}) = \\
&\sum_{t=1}^{2^q-1}\sum_{r=1}^{2^q-1} P\{\mathbf{0} \to \boldsymbol{t}\}P\{\mathbf{0} \to \boldsymbol{r}\}P\{\mathbf{0} \to \boldsymbol{t}B\}P\{\mathbf{0} \to \boldsymbol{t}A + \boldsymbol{r}B\}\cdot \\
&\qquad \cdot P\{\mathbf{0} \to \boldsymbol{r}A\}, \ \ \forall j, k \in \{0, 1, ..., 2^q - 1\}
\end{aligned} \tag{10}$$

Using (8) and (10) we derive that:

$$\begin{aligned}
P(A_2) = \sum_{t=1}^{2^q-1}\sum_{r=1}^{2^q-1} &P\{\mathbf{0} \to \boldsymbol{t}\}P\{\mathbf{0} \to \boldsymbol{r}\}P\{\mathbf{0} \to \boldsymbol{t}B\}\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{t}A + \boldsymbol{r}B\}P\{\mathbf{0} \to \boldsymbol{r}A\}
\end{aligned} \tag{11}$$

From (11) we see that $P(A_2)$ is independent from the true values of a_{i-1}, a_i, a_{i+1} and a_{i+2} , i.e., it is independent form the distribution of the characters in the input message. Obviously, $P(A_2)$ does not depend on the matrix C, too.

In general, to derive formula for $P(A_r)$ - the probability that exactly r consecutive characters $a_i, a_{i+1}, ...a_{i+r-1}$ from the input message are incorrectly transmitted and the error is not detected, we introduce random evens
$H_{j_0 j_1 ... j_{r-1}}$: the true value of a_i is j_0, the true value of a_{i+1} is j_1, the true value of a_{i+2} is j_2,..., the true value of a_{i+r-1} is j_{r-1}, where $j_0, j_1, ..., j_{r-1} \in \{0, 1, ..., 2^q - 1\}$.

Now,

$$P(A_r) = \sum_{j_0=0}^{2^q-1}\sum_{j_1=0}^{2^q-1} ... \sum_{j_{r-1}=0}^{2^q-1} P(A_r|H_{j_0 j_1 ... j_{r-1}})P(H_{j_0 j_1 ... j_{r-1}}) \tag{12}$$

$$P(A_r|H_{j_0 j_1 ... j_{r-1}}) = \sum_{\substack{s_0=0\\ s_0\neq j_0}}^{2^q-1}\sum_{\substack{s_1=0\\ s_1\neq j_1}}^{2^q-1} ... \sum_{\substack{s_{r-1}=0\\ s_{r-1}\neq j_{r-1}}}^{2^q-1} B_{s_0}^{s_{r-1}} \tag{13}$$

where in a same way as (9) we obtain:

$$\begin{aligned}
B_{s_0}^{s_{r-1}} = &P\{\mathbf{0} \to \boldsymbol{s_0} + \boldsymbol{j_0}\}P\{\mathbf{0} \to \boldsymbol{s_1} + \boldsymbol{j_1}\}P\{\mathbf{0} \to \boldsymbol{s_2} + \boldsymbol{j_2}\}\cdot ...\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{s_{r-1}} + \boldsymbol{j_{r-1}}\}P\{\mathbf{0} \to (\boldsymbol{s_0} + \boldsymbol{j_0})B\}\cdot \\
&\cdot P\{\mathbf{0} \to (\boldsymbol{s_0} + \boldsymbol{j_0})A + (\boldsymbol{s_1} + \boldsymbol{j_1})B\}\cdot \\
&\cdot P\{\mathbf{0} \to (\boldsymbol{s_1} + \boldsymbol{j_1})A + (\boldsymbol{s_2} + \boldsymbol{j_2})B\}\cdot ...\cdot \\
&\cdot P\{\mathbf{0} \to (\boldsymbol{s_{r-2}} + \boldsymbol{j_{r-2}})A + (\boldsymbol{s_{r-1}} + \boldsymbol{j_{r-1}})B\}\cdot \\
&\cdot P\{\mathbf{0} \to (\boldsymbol{s_{r-1}} + \boldsymbol{j_{r-1}})A\}
\end{aligned} \tag{14}$$

By introducing replacement $\boldsymbol{t_u} = \boldsymbol{s_u} + \boldsymbol{j_u}$, $u = 0, 1, 2, ..., r-1$ in the expression (14) and replacing it in (13), we get:

$$\begin{aligned}
P(A_r|H_{j_0 j_1 ... j_{r-1}}) = &\sum_{t_0=1}^{2^q-1}\sum_{t_1=1}^{2^q-1} ... \sum_{t_{r-1}=1}^{2^q-1} P\{\mathbf{0} \to \boldsymbol{t_0}\}\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{t_1}\}P\{\mathbf{0} \to \boldsymbol{t_2}\}\cdot ...\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{t_{r-2}}\}P\{\mathbf{0} \to \boldsymbol{t_{r-1}}\}\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{t_0}B\}P\{\mathbf{0} \to \boldsymbol{t_0}A + \boldsymbol{t_1}B\}\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{t_1}A + \boldsymbol{t_2}B\}\cdot ...\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{t_{r-2}}A + \boldsymbol{t_{r-1}}B\}\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{t_{r-1}}A\}
\end{aligned} \tag{15}$$

From (12) and (15), we derive

$$\begin{aligned}
P(A_r) = &\sum_{t_0=1}^{2^q-1}\sum_{t_1=1}^{2^q-1} ... \sum_{t_{r-1}=1}^{2^q-1} P\{\mathbf{0} \to \boldsymbol{t_0}\}\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{t_1}\}P\{\mathbf{0} \to \boldsymbol{t_2}\}\cdot ...\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{t_{r-2}}\}P\{\mathbf{0} \to \boldsymbol{t_{r-1}}\}\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{t_0}B\}P\{\mathbf{0} \to \boldsymbol{t_0}A + \boldsymbol{t_1}B\}\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{t_1}A + \boldsymbol{t_2}B\}\cdot ...\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{t_{r-2}}A + \boldsymbol{t_{r-1}}B\}\cdot \\
&\cdot P\{\mathbf{0} \to \boldsymbol{t_{r-1}}A\}
\end{aligned} \tag{16}$$

This means that $P(A_r)$ is independent from the distribution of the characters in the input message and from the matrix C. Thus, the theorem is proven. ■

Using the fact that the probability of undetected errors is independent from the distribution of the characters in the input message the following theorem holds (proved in [16]):

Theorem 2: Let $f(n, p)$ be the probability that at most 4 characters of the input message with length n are incorrectly transmitted through a binary symmetric channel with probability of bit-error p and the error is not detected. If linear

quasigroup of order 2^q is used for the code defined with (3) then

$$\begin{aligned}
f(2,p) &= 2v_0v_1 + r_2\\
f(3,p) &= 3v_0^3v_1 + 3v_0v_2 + r_3\\
f(4,p) &= 4v_0^5v_1 + 4v_0^3v_2 + 2v_0^2v_1^2 + 4v_0v_3 + r_4\\
f(n,p) &= nv_1v_0^{2n-3} + nv_2v_0^{2n-5} + \frac{n(n-3)}{2}v_1^2v_0^{2n-6}\\
&\quad + nv_3v_0^{2n-7} + n(n-4)v_2v_1v_0^{2n-8}\\
&\quad + \frac{n(n-4)(n-5)}{6}v_1^3v_0^{2n-9} + nv_4v_0^{2n-9}\\
&\quad + n(n-5)v_3v_1v_0^{2n-10} + \frac{n(n-5)}{2}v_2^2v_0^{2n-10}\\
&\quad + \frac{n(n-5)(n-6)}{2}v_2v_1^2v_0^{2n-11}\\
&\quad + \frac{n(n-5)(n-6)(n-7)}{24}v_1^4v_0^{2n-12},\ \ n \geq 5
\end{aligned} \tag{17}$$

In the formulas, we use the following notations:
v_t - the probability of undetected errors when exactly t consecutive characters of the initial message $a_0a_1 \dots a_{n-1}$ are incorrectly transmitted (the characters $a_i, a_{i+1}, \dots, a_{i+t-1}$ are incorrectly transmitted, but a_{i-1} and a_{i+t} are correctly transmitted), $t = 1, 2, 3, 4$;
v_0 - the probability of correct transmission of a character;
r_t - the probability of undetected errors in a block with length t when all t characters are incorrectly transmitted, $t = 2, 3, 4$.

Although the Theorem 2 in [16] is formulated for fractal quasigroups of order 4, from the proof it can be seen that it holds if for coding is used quasigroup of arbitrary order for which the probability of undetected errors is independent from the distribution of the characters in the input message.

The formula (17) gives us an approximate formula for the probability of undetected errors. Namely, the probability that 5 or more characters of the input message are incorrectly transmitted and the error is not detected is inconsiderably small for small values of a bit-error p. For this reason and the fact that in the real channels the probability of bit-error p is very small, the formula $f(n,p)$ given with (17) gives a good enough approximation of the probability of undetected errors.

The parameters v_t in Theorem 2 are practically $P(A_t)$ from the proof of Theorem 1. The parameters r_2, r_3 and r_4 occur for the following reason. Let say that the two consecutive characters a_0 and a_1 are incorrectly transmitted. The information character a_0 affects the redundant characters d_{n-1} and d_0, while a_1 affects d_0 and d_1. If the block length is greater then or equal to 3, then a_0 and a_1 have one common redundant characters and both of them affect d_0. But if the block length is equal to 2, then the characters a_0 and a_1 have two common redundant characters that are affected: d_0 and d_1. For this reason, the probability that two consecutive characters are incorrectly transmitted and the error is not detected for the blocks with length two is different than the probability for the blocks with length greater than two. Therefore, this case should be considered separately from the general one, and requesting the parameter r_2 to be introduced. A similar situation is for r_3 and r_4. The formulas for these parameters are obtained analogously to the formulas for v_t:

$$\begin{aligned}
r_2 = \sum_{t=1}^{2^q-1}\sum_{r=1}^{2^q-1} & P\{\mathbf{0}\to\boldsymbol{t}\}P\{\mathbf{0}\to\boldsymbol{r}\}P\{\mathbf{0}\to\boldsymbol{t}A+\boldsymbol{r}B\}\cdot\\
&\cdot P\{\mathbf{0}\to\boldsymbol{r}A+\boldsymbol{t}B\}
\end{aligned} \tag{18}$$

$$\begin{aligned}
r_3 = \sum_{t=1}^{2^q-1}\sum_{r=1}^{2^q-1}\sum_{s=1}^{2^q-1} & P\{\mathbf{0}\to\boldsymbol{t}\}P\{\mathbf{0}\to\boldsymbol{r}\}P\{\mathbf{0}\to\boldsymbol{s}\}\cdot\\
&\cdot P\{\mathbf{0}\to\boldsymbol{t}A+\boldsymbol{r}B\}P\{\mathbf{0}\to\boldsymbol{r}A+\boldsymbol{s}B\}\cdot\\
&\cdot P\{\mathbf{0}\to\boldsymbol{s}A+\boldsymbol{t}B\}
\end{aligned} \tag{19}$$

$$\begin{aligned}
r_4 = \sum_{t=1}^{2^q-1}\sum_{r=1}^{2^q-1}\sum_{s=1}^{2^q-1}\sum_{h=1}^{2^q-1} & P\{\mathbf{0}\to\boldsymbol{t}\}P\{\mathbf{0}\to\boldsymbol{r}\}P\{\mathbf{0}\to\boldsymbol{s}\}\cdot\\
&\cdot P\{\mathbf{0}\to\boldsymbol{h}\}P\{\mathbf{0}\to\boldsymbol{t}A+\boldsymbol{r}B\}\cdot\\
&\cdot P\{\mathbf{0}\to\boldsymbol{r}A+\boldsymbol{s}B\}P\{\mathbf{0}\to\boldsymbol{s}A+\boldsymbol{h}B\}\cdot\\
&\cdot P\{\mathbf{0}\to\boldsymbol{h}A+\boldsymbol{t}B\}
\end{aligned} \tag{20}$$

In order to calculate the probability of undetected errors for a given linear quasigroup, one should first calculate the values of the parameters, using (16), (18), (19) and (20), and then to substitute these values into (17).

V. Results With Linear Quasigroups of Order 8

A. The Smallest Probability of Undetected Errors

The values of v_t depend on matrices A and B (see (16)), from where it follows that they depend on the chosen quasigoup for coding. Since the probability of undetected errors depends on v_t, it follows that this probability depends on the chosen quasigroup for coding. It is best the probability of undetected errors to be as small as possible. For this reason, we applied the formula (17) on each pair (A, B) of non-singular binary matrices of order 3 and found that the smallest probability of undetected errors is the following:

$$\begin{aligned}
f(2,p) &= (1-p)^2p^3(4 - 20p + 56p^2 - 96p^3 + 96p^4\\
&\quad - 55p^5 + 22p^6 - 3p^7)\\
f(3,p) &= (1-p)^3p^3(3 - 30p + 162p^2 - 580p^3 + 1470p^4\\
&\quad - 2658p^5 + 3394p^6 - 3024p^7 + 1866p^8 - 787p^9\\
&\quad + 213p^{10} - 27p^{11} + p^{12})\\
f(4,p) &= (1-p)^4p^4(8 - 88p + 404p^2 - 784p^3 - 808p^4\\
&\quad + 9440p^5 - 29720p^6 + 57432p^7 - 77044p^8\\
&\quad + 74352p^9 - 51892p^{10} + 25960p^{11} - 9179p^{12}\\
&\quad + 2268p^{13} - 378p^{14} + 44p^{15} - 3p^{16})\\
f(n,p) &= \tfrac{1}{24}np^4(1-p)^{6(n-4)} \times \Big[24 - 384p + 2832p^2\\
&\quad - 12384p^3 + 12(n + 2807)p^4 - 48(2n + 1009)p^5\\
&\quad + 48(6n - 499)p^6 + 326976p^7 + 4(n^2 - 603n\\
&\quad - 239092)p^8 + 48(144n + 37283)p^9 - 24(n^2\\
&\quad + 271n + 102896)p^{10} + 24(5n^2 - 455n\\
&\quad + 110576)p^{11} + (n^3 + 78n^2 + 39863n\\
&\quad - 2273238)p^{12} + 8(n^3 - 96n^2 - 6943n\\
&\quad + 194358)p^{13} + 4(5n^3 - 16n^2 + 13801n\\
&\quad - 213770)p^{14} + 8(n^3 + 99n^2 - 4924n + 46854)p^{15}\\
&\quad - 2(13n^3 + 420n^2 - 10207n + 64386)p^{16}\\
&\quad - 8(n^3 - 86n^2 + 1007n - 4234)p^{17} + 4(5n^3\\
&\quad - 90n^2 + 583n - 1626)p^{18} - 8(n^3 - 12n^2 + 53n\\
&\quad - 102)p^{19} + (n^3 - 10n^2 + 35n - 50)p^{20}\Big], n \geq 5
\end{aligned} \tag{21}$$

The graphic of this function, for different values of the block length n is given in Figure 2.

B. Controlling the Error

As we can see from Figure 2, when the block length increases the probability of undetected errors decreases and the sequence of maximums converges to zero. This means that there is some natural number n_0, such that the maximum of $f(n,p)$ will be smaller than ε for all natural numbers n that are greater than or equal to n_0 and the maximum of $f(n,p)$ will be greater than ε for all natural numbers n that are smaller than n_0. So, if we want the probability of undetected errors to be smaller than some previous given value ε, we will choose

 ISBN: 978-1-61208-411-4

the block length n to be the smallest natural number such that the maximum of the function $f(n,p)$ is smaller than ε (i.e., $n = n_0$). Since the maximum of $f(n,p)$ will be smaller than ε, follows that $f(n,p)$ will be smaller than ε for all values of $p \in (0, \frac{1}{2})$. We choose n to be the smallest natural number for which the maximum of the function $f(n,p)$ is smaller than ε since we want if there are errors in transmission the smallest possible blocks to be retransmitted.

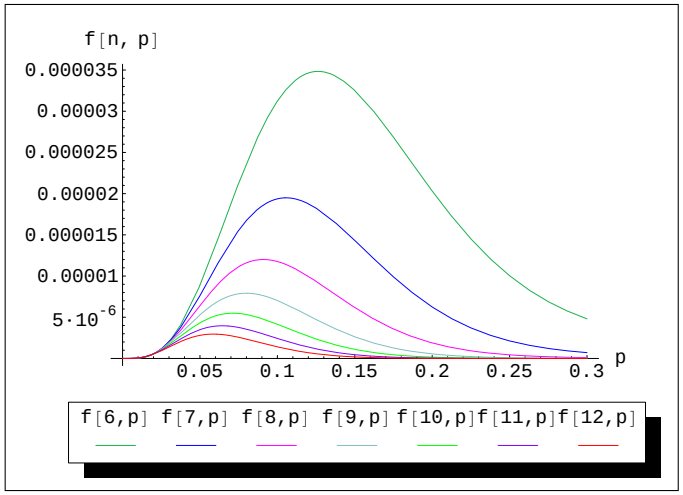


Figure 2. The smallest probability of undetected errors for different values of the block length n if for coding are used quasigroups of order 8.

C. The Best Class of Linear Quasigroups of Order 8

We define the best class of linear quasigroups of order 8 to be the class that contains exactly those linear quasigroups of order 8 that achieve the smallest probability of undetected errors, i.e., the probability given with (21). The quasigroups from this class are best for coding.

We obtained that the best class contains the quasigroups for which the matrix A contains exactly two zeros and the matrix B is determined by the matrix A in the following way. For each non-singular binary matrix A of order 3 with exactly 2 zeros, there are two possible choices for the matrix B: one in which the rows of A are cyclically shifted one position up, and the other in which the rows of A are cyclically shifted two positions up. Namely, if $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ is non-singular binary matrix with exactly two zeros, than this matrix A determine two quasigroups of order 8 which have the smallest probability of undetected errors. One of them is obtained when $B = \begin{bmatrix} a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{11} & a_{12} & a_{13} \end{bmatrix}$ and the other is obtained when $B = \begin{bmatrix} a_{31} & a_{32} & a_{33} \\ a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$.

There are 18 binary non-singular matrices with two zeros that can be chosen for the matrix A and for each one of them there are two possible choice for the matrix B. Therefore, there are 36 pairs of non-singular binary matrices (A, B) which give the smallest probability of undetected errors. These quasigroups are given below. Since for every of these 36 pairs of matrices (A, B), any of the eight binary matrices of order 1×3 may be chosen as matrix C, there are 288 linear quasigroups of order 8 in the best class of linear quasigroups of order 8.

The 36 pairs of non-singular binary matrices (A, B) in the best class are given in Figure 3.

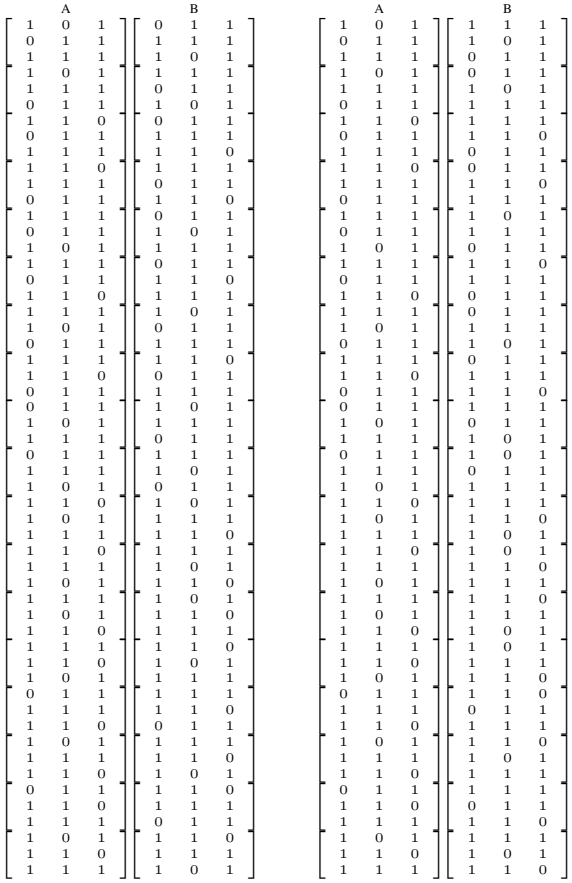


Figure 3. The best class of linear quasigroups of order 8.

VI. Comparison With the CRC Codes

We give a comparison of our codes with some CRC codes. The comparison is made from the aspect of the probability of undetected errors. Ability of the CRC code to detect errors depends on the chosen polynomial for coding. We will consider several cases of polynomials, accepted as a standard for coding. Namely, we will consider CRC-12 defined with the polynomial $g(x) = x^{12}+x^{11}+x^3+x^2+x+1$ [4], CRC-ANSI defined with $g(x) = x^{16} + x^{15} + x^2 + x + 1$ [19] and CRC-CCITT defined with $g(x) = x^{16} + x^{12} + x^5 + 1$ [20]. Unlike our code, CRC code adds constant number of redundant characters, regardless of the length of the information block. CRC-12 adds 12 bits, while CRC-ANSI and CRC-CCITT add 16 bits. In Table I the maximums of the probabilities of undetected errors for our code and CRC-12 code are given, while in Table II the maximums of the probabilities of undetected errors for our code, CRC-ANSI and CRC-CCITT codes for different values of the block length n are given. The values for CRC codes are taken from [18]. The block length is expressed in bits. We can see that our code has smaller probability of undetected errors than CRC-12 for all values of block length n which are greater than or equal to 12. For n greater than or equal to 15, our code has smaller probability of undetected errors than all considered

 ISBN: 978-1-61208-411-4

TABLE I. THE MAXIMUMS OF THE PROBABILITY OF UNDETECTED ERRORS FOR OUR CODE AND CRC-12. THE BLOCK LENGTH n IS EXPRESSED IN BITS

n	Our code	CRC-12
6	1.53809×10^{-2}	4.98239×10^{-4}
9	1.94931×10^{-3}	4.44429×10^{-4}
12	2.72046×10^{-4}	4.92904×10^{-4}
15	7.22452×10^{-5}	5.32493×10^{-4}
18	3.48346×10^{-5}	5.24102×10^{-4}
21	1.94926×10^{-5}	5.01021×10^{-4}
24	1.20057×10^{-5}	4.95408×10^{-4}
27	7.91486×10^{-6}	4.68967×10^{-4}
30	5.49112×10^{-6}	4.36575×10^{-4}
33	3.96425×10^{-6}	4.13106×10^{-4}
36	2.95482×10^{-6}	3.98553×10^{-4}
39	2.26092×10^{-6}	3.84785×10^{-4}
42	1.76831×10^{-6}	3.68207×10^{-4}
45	1.40897×10^{-6}	3.53279×10^{-4}
48	1.14074×10^{-6}	3.40974×10^{-4}

TABLE II. THE MAXIMUMS OF THE PROBABILITY OF UNDETECTED ERRORS FOR OUR CODE, CRC-ANSI AND CRC-CCITT. THE BLOCK LENGTH n IS EXPRESSED IN BITS

n	Our code	CRC-ANSI	CRC-CCITT
6	1.53809×10^{-2}	2.09564×10^{-4}	1.82571×10^{-4}
9	1.94931×10^{-3}	1.83062×10^{-4}	1.59587×10^{-4}
12	2.72046×10^{-4}	1.49497×10^{-4}	1.31108×10^{-4}
15	7.22452×10^{-5}	1.49435×10^{-4}	1.07281×10^{-4}
18	3.48346×10^{-5}	1.87672×10^{-4}	9.68045×10^{-5}
21	1.94926×10^{-5}	1.96955×10^{-4}	8.80828×10^{-5}
24	1.20057×10^{-5}	1.88110×10^{-4}	7.82445×10^{-5}
27	7.91486×10^{-6}	1.72350×10^{-4}	6.89410×10^{-5}
30	5.49112×10^{-6}	1.66609×10^{-4}	6.05393×10^{-5}
33	3.96425×10^{-6}	1.67740×10^{-4}	5.32930×10^{-5}
36	2.95482×10^{-6}	1.61975×10^{-4}	4.71277×10^{-5}
39	2.26092×10^{-6}	1.52149×10^{-4}	4.18904×10^{-5}
42	1.76831×10^{-6}	1.40941×10^{-4}	3.74422×10^{-5}
45	1.40897×10^{-6}	1.34158×10^{-4}	3.41088×10^{-5}
48	1.14074×10^{-6}	1.31914×10^{-4}	3.17809×10^{-5}

CRC codes (even in the case when the CRC checksum is for short lengths such that the CRC code has also a rate of 1/2).

Additionally, which is important for Safety-Crytical Systems, we can make the probability of undetected errors arbitrary small, which is not case with CRC codes. Namely, the probability of undetected errors for CRC code with c redundant bits tends to 2^{-c} when the block length n tends to infinity.

VII. CONCLUSION

We defined error-detecting codes based on linear quasigroups. We proved that the probability of undetected errors is independent from the distribution of the characters in the input message. Using this property, we found the best class of linear quasigroups of order 8 for such coding and we computed the corresponding probability of undetected errors. Finally, we explained how the probability of undetected errors can be made arbitrary small. We compare our codes with CRC-12, CRC-ANSI and CRC-CCITT and show that our code has smaller probability of undetected errors than the CRC codes when code rate and block lengths are equal.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss.Cyril and Methodius" University.

REFERENCES

[1] W. W. Peterson and D. T. Brown, "Cyclic Codes for Error Detection," in Proceedings of the IRE, vol. 49, no. 1, 1961, pp. 228-235.

[2] P. Koopman and T. Chakravarty, "Cyclic Redundancy Code (CRC) Polynomial Selection For Embedded Networks," International Conference on Dependable Systems and Networks, 2004, pp. 145-154.

[3] A. Richardson, WCDMA Handbook, Cambridge, UK, Cambridge University Press, 2005, pp. 223.

[4] A. Perez, "Byte-Wise CRC Calculations," IEEE Micro vol. 3, no. 3, 1983, pp. 40-50.

[5] S. Blanc, A. Bonastre, and P. J. Gil, "Dependability assessment of by-wire control systems using fault injection," Journal of Systems Architecture, vol. 55, no. 2, 2009, pp. 102-113.

[6] P. Koopman, "32-Bit Cyclic Redundancy Codes for Internet Applications," The International Conference on Dependable Systems and Networks, 2002, pp. 459-468.

[7] D. T. Jones, "An improved 64-bit cyclic redundancy check for protein sequences," University College London, 2009.

[8] J. G. Fletcher, "An Arithmetic Checksum for Serial Transmissions," IEEE Transactions on Communications, vol 30, no. 1, 1982, pp. 247-252.

[9] J. Zweig and C. Partridge, "TCP Alternate Checksum Options," IETF RFC 1146, Mar. 1990.

[10] P. Deutsch and J.-L. Gailly, "ZLIB Compressed Data Format Specification Version 3.3," IETF RFC 1950, May 1996.

[11] G. Latif-Shabgahi, J. M. Bass, and S. Bennett, "A taxonomy for software voting algorithms used in safety-critical systems," IEEE Transactions on Reliability, vol. 53, no. 3, 2004, pp. 319-328.

[12] V. Bakeva and N. Ilievska, "A probabilistic model of error-detecting codes based on quasigroups," Quasigroups and Related Systems, vol. 17, no. 2, 2009, pp. 135-148.

[13] N. Ilievska and D. Gligoroski, "Error-Detecting Code using Linear Quasigroups," Advances in Intelligent Systems and Computing vol. 311, ICT Innovations 2014, Springer, 2014, pp. 309-318.

[14] N. Ilievska and D. Gligoroski, "An Error-Detecting Code based on Linear Quasigroups," in Proceedings of 11th International Conference for Informatics and Information Technology (CIIT 2014), Bitola, Republic of Macedonia, 2014, in press.

[15] N. Ilievska, "Proving the probability of undetected errors for an error-detecting code based on quasigroups," Quasigroups and Related Systems vol. 22, no. 2, 2014, pp. 223-246.

[16] N. Ilievska and V. Bakeva, "A Model of error-detecting codes based on quasigroups of order 4," in Proceedings of 6th International Conference for Informatics and Information Technology, Bitola, Republic of Macedonia, 2008, pp. 7-11.

[17] Y. Chen, M. Niemenmma, A.J. Han Vinck, and D. Gligoroski, "On the Error Detection Capability of One Check Digit," IEEE Transactions on Information theory, 2014, pp. 261-270.

[18] K. A. Witzke, "Exmanination of the undetected error probability of linear block codes," Thesis: M.A. Sc, University of British Columbia Department of Electrical Engineering, 1984.

[19] T.V. Ramabadran and S.S. Gaitonde, "A tutorial on CRC computations," Micro, IEEE, vol.8, no.4, Aug. 1988, pp. 62-75.

[20] P. Koopman and T. Chakravarty, "Cyclic Redundancy Code (CRC) Polynomial Selection For Embedded Networks," in Proceedings of the International Conference on Dependable Systems and Networks, 2004, pp. 145-154.

[21] J.C. Knight, "Safety Critical Systems: Challenges and Directions," in Proceedings of the 24th International Conference on Software Engineering, 2002, pp. 547-550.

 ISBN: 978-1-61208-411-4

Radio Access Scheme using Super Pilot Channel in Reconfigurable Multi RAT-based Wireless Communication System

Woogoo Park
Wired & Wireless Convergence Research Department
ETRI
Daejeon, KOREA
wgpark@etri.re.kr

Hoyoung Song
Wired & Wireless Convergence Research Department
ETRI
Daejeon, KOREA
hsong@etri.re.kr

***Abstract*—In this paper, we propose a new radio access method using super pilot channel in reconfigurable multi Radio Access Technology (RAT)-based wireless communication system. The goals of the proposed method are directed to a system and a process for radio access having compatibility with existing systems, being capable of increasing frequency efficiency, and being capable of increasing the transmitting rate. The intermediate results of the paper lay the ground for designing a new 5G air interface beyond LTE-A, which suits the diverse needs of future applications, like interference coordination between small cells and macro cells.**

Keywords-reconfigurability; wireless communication; pilot channel; super pilot channel; micro-band; macro-band; dynamic spectrum allocation; flexible spectrum management

I. INTRODUCTION

Various standards for wireless communication technologies have been established. Global wireless network operators have been taking steps to advance throughput of their mobile networks as part of the fourth generation LTE (Long-Term Evolution) communications technology. An important factor in radio access is the interference among cells [1]. By a very dense deployment of low-cost, low-power base stations, both the spatial reuse of radio resource and transmit power efficiency can be potentially improved. It is envisioned that the next generation wireless networks will consist of macro-cells and a high density of small-cells with different capabilities including transmit power and coverage range [2]. Also, to improve spectrum efficiency, the D2D (Device-to-Device) communication is one of the solutions in heterogeneous networks [3]. The interference between macro-cells and small-cells as well as the interference between adjacent small cells is always a serious concern. The widely used inter-cell interference (ICI) mitigation techniques in homogeneous networks are soft frequency reuse [1] and interference self-cancellation scheme [4]. Although small cells can help reduce data traffic density, complexity should be improved in using ICI. Especially ICI mitigation techniques in D2D should be applied to both small and wide area cells; to both low and high frequency bands; to both high and low mobility scenarios; and also it could improve the effective SINR (Signal-to-Interference and Noise Ratio). The core concept of the fourth generation network is as follows: every (Signal-to-Interference and Noise Ratio). The core concept of the fourth generation network is as follows: every device uses an IP address, and the proposed network of a convergence type includes an IP-based core network and access networks based on various existing standards. The fact that these various standards operate in different bands restricts any approach to accommodate all future standards. In most cases, existing sensing frequency bands have a too wide range of 400MHz to 6GHz [5]. Therefore, it takes a long sensing time to use a different system, and a large amount of power is consumed. The advantage of this paper is that it allows tight coordination features such as interference in D2D. It also provides potential for spectrum gains like easier deployment and other site cost. The rest of this paper is organized as follows: Section II describes the related works. In Section III, the system description and the proposed frequency structure of a cell are presented. Then in Section IV, the radio access scheme with access method of macro-band and micro-band SPC (Super Pilot Channel) in RAS is considered. Conclusion is shown in Section V.

II. RELATED WORKS

Small cells in heterogeneous network are typically overlaid on the existing macro cells and installed in the dense area close to small cell users [2]. It is noted that small cell users need sensing time for avoiding interference from macro cell. Compared to the interference management of D2D communication on the different frequencies in the previous work [3], we propose interference management not only D2D but also general mobile users by allocating SPC-based channel over macro cell area including small cells. In an attempt to reduce sensing time among heterogeneous networks, E^2R is currently developing concepts and solutions for a Cognitive Pilot Channel (CPC), encompassing both in-band/out-band and downlink/uplink functionalities [6]. It focuses on the network selection strategy according to the information which could be brought by CPC, whereas our key ideas are reconfigurable and broadcast-based SPC assignment by designing of frequency structure of macro and micro cells. Another works is that a novel homogeneous mesh grouping scheme based broadcast CPC mode is designed to improve the efficiency of broadcast CPC mode in the Cognitive Wireless Networks

 ISBN: 978-1-61208-411-4

[7]. However the CPC is based on broadcasting technology and each access station is provided with multiple RAT information. This requires a supervising CPC control station for managing the RAT overall. J. R. Moorman presented the development and implementation of a software radio designed for a 3G system that expands upon the notion of the physical layer software radio to encompass upper layer processing capabilities [8].

To realize the structure of device with a high mobility, flexibility and reconfigurability, software Defined Radio (SDR) is one of possibilities. It provides the seamless shifting between existed air-interface standards. Extending the flexibility further, a system capable to sense the spectrum space available for communication and adapt to it is Cognitive Radio (CR). Obviously SDR in CR should be configured not only to independent standards, protocols and services but also to the extensively dynamic nature of bandwidth allocation [9]. In [10], a low-cost reconfigurable antenna array was implemented for SDR-based Communication Systems.

In this paper, we propose a prototype of a reconfigurable radio access for microcells as well as macro cell, based on SPC that will be integrated with multi-RAT networks and shown operation of a micro-band based on a micro-band SPC between a RAS and its RMS. SPC is used through interworking with a Reconfigurable Mobile Station (RMS) to provide an optimum radio access environment satisfying Dynamic Spectrum Allocation (DSA) and Flexible Spectrum Management (FSM). The first RAS, which evolves to enable SPC-based multi-access, shares corresponding SPC information through broadcasting (macro-band) to support radio environment sharing and reconfiguration of the RMS based on the sharing (micro- band). As used herein, "reconfigurable" means that a number of RATs are supported, and RMSs can be configured in conformity with each RAT. Such technology includes CR/SDR technology, etc. In practice, a large number of CPC control stations are expectably necessary on a global scale, and a considerable amount of cost and time will be incurred. In this paper, we focus on inter-cell interference between the macro cell and the small cell as well as frequencies allocation under heterogeneous networks.

III. PROPOSED MODEL

A system for radio access in a wireless communication system in which a number of RATs exist includes RAS configured to share radio environment information and the RATs with an adjacent RAS using a macro-band SPC. The RAS being reconfigurable in conformity with RATs and a RMS configured to transmit and receive the radio environment information and the RATs to/from the RAS using a micro-band SPC and access the RAS using the micro-band SPC, the RMS being reconfigurable in conformity with RAT of the accessed RAS.

A. System Description

Fig. 1 briefly depicts a method for radio access based on a SPC in a reconfigurable multi-RAT mobile communication system. It shows RASs in RAT-i to handle RATs supported in respective cells. RMSs in RAT-j access the RASs and receive a service using the RATs. The macro-band SPCs (red arrows) in RAT-1 exchange information regarding the RATs between the RASs. The micro-band SPCs (blue arrows) in RAT-j perform access and control between the RASs and the RMSs. Each RAS supports RAT, which is supported by a cell, managed by the RAS itself. Respective RASs configured to transmit the macro-band SPCs may be configured in a mesh type. Various RATs are used in Fig. 1 and RATs have different cell radius. In the overlapping cell environment using RAT-1, respective RASs belonging to RATs (RAT-2, RAT-3, RAT-4, RAT-i, and RAT-j) use their own RATs. In a cell using RAT-2, the current RAT-1 can be used simultaneously (i.e. overlapping cell). For example, a WiBro cell capable of managing a wide range of networks may include a WLAN cell capable of managing small-scale networks. A RMS in RAT-j transmits an access request to an accessible RAM in RAT-i using micro-band SPC between RAS-j and RAM in RAT-j and, when the access request is acknowledged, communication becomes possible. In order to enable this process, the RAS in the center cell broadcasts RAT and radio environment information to RASs in adjacent cells using a macro-band SPC between center RAS and RAS-j in RAT-j to share the radio environment information. Based on the information broadcasted using the macro-band SPCs between RAS-j and RAM in RAT-j, when the access is acknowledged, communication becomes possible.

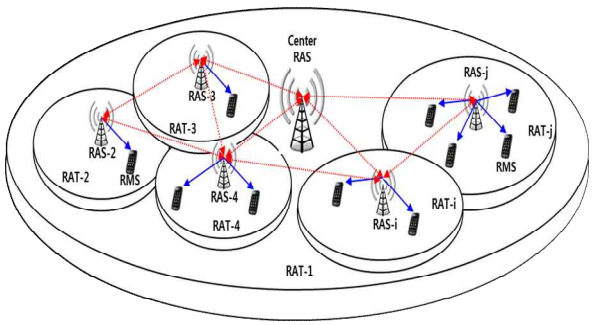


Figure 1. Super pilot channel-based cell configuration

B. Arrangement of Frequencies

In Fig. 2, single center cell is surrounded by six adjacent cells and it communicates with RMSs using same frequency. For example, a cell managed by RAS1 is surrounded by six adjacent cells managed by six RASs{2,3,4,5,6,7}. Furthermore, a cell managed by RAS3 is surrounded by six adjacent cells managed by six RASs{1,2,4,8,9,10}. The RASs of such center cells broadcast radio environment information using the

macro-band SPCs to share the radio environment information. For example, RAS3 is an adjacent cell of RAS1 and receives radio environment information through the macro-band SPC.

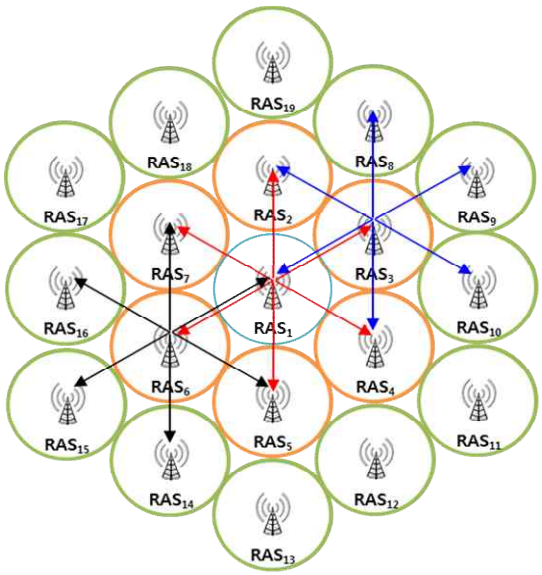


Figure 2. Arrangement of frequencies among RASs

However when RAS3 acts as a center cell, it broadcasts radio environment information to RASs of its six adjacent cells through the macro-band SPC to share the radio environment information. Each unit cell has a first layer of cells, where influence is limited to an adjacent cell by adjusting power intensity without using different frequencies, and a second layer of cells, where a single frequency band is used to communicate with adjacent cells to avoid interference with cells beyond the adjacent cells. For example, the first layer of RAS_1 includes $RASs_{\{12,3,4,5,6,7\}}$ and second layer thereof includes $RASs_{\{8,9,10\}}$.

C. Frequency Structure of a Cell

Fig. 3 shows a RAS with an antenna belonging to the RAS and capable of transmitting a micro-band SPC and a macro-band SPC. The i^{th} RAS has two antennas and two frequency bands corresponding to the two antennas. One frequency band is used to broadcast radio environment information to the $j^{th}(j>i)$ RAS in an adjacent cell using the antenna which belongs to the RAS and which can transmit a macro-band SPC.

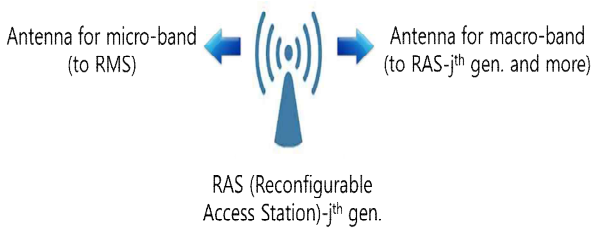


Figure 3. Frequency structure of RAS

In addition, RMSs inside the i^{th} RAS are provided with radio environment information regarding the center cell and adjacent cells using the antenna, which can transmit a micro-band SPC. The two antennas are configured to transmit/receive two different frequency bands respectively, i.e. a macro-band as a frequency band for broadcasting each radio environment information to the $j^{th}(j>i)$ RAS in an adjacent cell and a micro-band as a frequency band for providing RMSs inside the i^{th} RAS with radio environment information regarding the center cell and adjacent cells.

IV. Radio Access Scheme

To cope with the huge demand for capacity in ultra-dense network, next-generation networks rely on densely deployed RAS between macro and small cells. To expand capacity and minimize interference, we used macro-band SPC on between RASs and micro-band between RAS and its RMSs.

A. Access Method of Macro-band & Micro-band SPC in RAS

Figs. 4 and 5 show access method of macro-band and micro-band SPC, which transmits and receives radio environment information between RASs. The RAS in Fig. 4 transmits radio environment information, which has been measured and stored in its storage space, to adjacent RASs including the RAS-2 using the macro-band SPC. The RAS-2 stores the radio environment from the RAS-1 in its storage space, acting as center RAS, and transmits the information from the RAS-1 to RASs including the RAS-3 using the macro-band SPC. In a similar manner, the RAS-3 receives the radio environment information from the RAS-1 and stores the information in its storage space. The RAS-2 similarly transmits its radio environment information to RASs in adjacent cells, i.e. RAS-1 and RAS-3, which then stores the radio environment information from the RAS-2, acting as center RAS, and transmits radio environment information regarding the RAS-1 to RASs in adjacent cells.

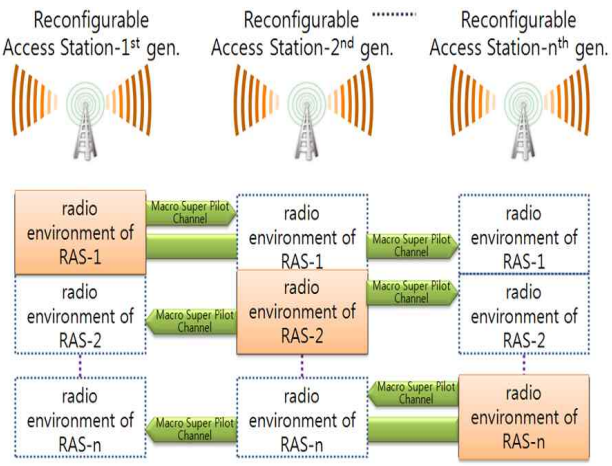


Figure 4. Access method of macro-band SPC between RASs

Fig. 5 shows operation of a micro-band based on a micro-band SPC between a RAS and its RMS and also represents a process of accessing RAS-*i* by RMS using micro-band SPCs (REQ & ACK). It is assumed that RAS-1 supports RAT-1.

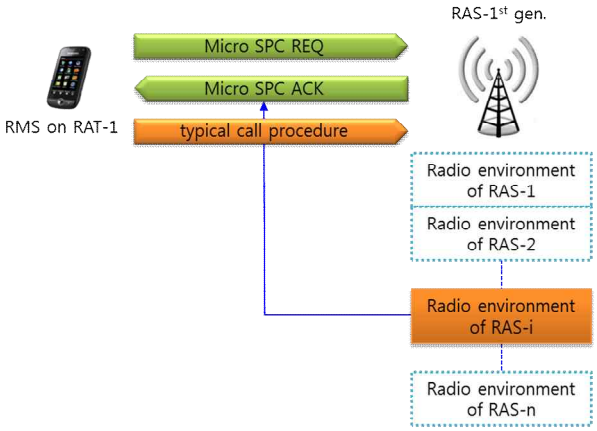


Figure 5. Operation of micro-band SPC between RAS and its RMS

B. Handover Mchanism in SPC Operation

It is assumed that in Fig. 6, RAT-1 is supported by RAS-i, and RAT-2 is supported by RAS-*j*. The RMS existing inside a cell of RAT-1 transmits a micro-band SPC REQ message, which is an in-band signal, to the RAS-*i*.

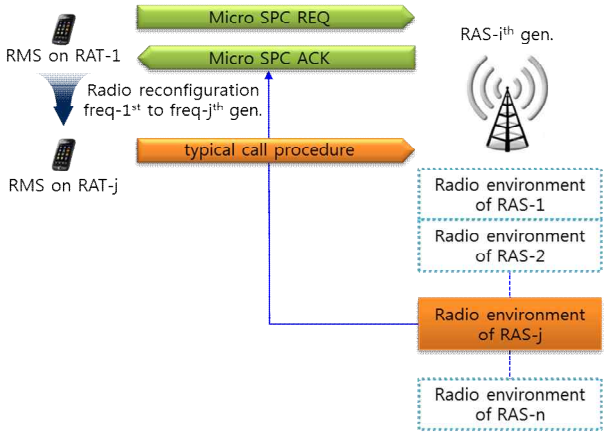


Figure 6. Operation of handover by exchanging adjacent channel information

The RAS-i transmits its radio environment information as shown in Fig. 5, when the radio environment is available. However, when the radio environment is not available, the RAS-i checks the radio environment information of an adjacent cell to see which is more available. Then the RAS-i loads frequency of an adjacent cell, which is the most available, onto a micro-band SPC ACK message and transmits it to the RMS that has made the request. The RMS receives the frequency regarding the RAS-j and changes it into RAT. The RMS moves to the RAS-j and performs a typical call procedure. In Table I, we provide an overview of the structure of radio environment information believed to be closely related to the radio environment information frame of a RAS considered in this paper. Each cell has its radio environment information map data including a self-RAS id field containing its own RAS id and an adjacent RAS id field containing information regarding operators to which RASs belong. In addition, the frequency, radio access specification, channel status, and traffic status are stored for respective operators to which RASs belong. Fig. 7 illustrates the structure of a macro-band SPC.

TABLE I. RADIO ENVIRONMENT INFORMATION FRAME

Self RAS id	Adjacent RAS is	Operator	Frequency	RAT	Channel Status	Traffic
RAS_1	RAS_1	O_1	f_1	RAT_1	ch_sts	traffic_sts
			f_2	RAT_2	ch_sts	traffic_sts
		O_2	f_3	RAT_3	ch_sts	traffic_sts
			f_2	RAT_2	ch_sts	traffic_sts
	RAS_2	O_1	f_1	RAT_1	ch_sts	traffic_sts
			f_2	RAT_2	ch_sts	traffic_sts
		O_2	f_3	RAT_3	ch_sts	traffic_sts
			f_2	RAT_2	ch_sts	traffic_sts

A macro-band is similar to an out-of-band signal of a CPC. And radio environment information between RASs includes a REQ message and an ACK message. The REQ message includes a RAS id field and a null field. The ACK message of the receiving RAS corresponds to an ACK signal in response to the REQ signal of the transmitting RAS, and includes self-RAS id and its radio environment information. Fig. 7 illustrates an overview of a macro-band SPC, which is similar to an out-of-band signal of a CPC. It refers to a signal for sharing RAS radio environment information between RASs, and includes an RAS and its radio environment information transmitted between RASs. Fig. 7 illustrates an overview of a macro-band SPC, which is similar to an out-of-band signal of a CPC. It refers to a signal for sharing RAS radio environment information between RASs, and includes an RAS and its radio environment information transmitted between RASs.

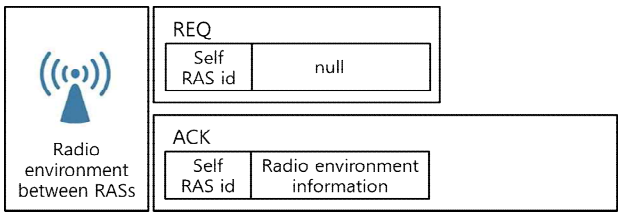


Figure 7. Message of macro-band SPC

The radio environment information includes a REQ message and an ACK message. The REQ message consists of a RAS id and a null. The ACK message consists of RAS id and radio environment information. The RAS REQ message of a transmitting RAS corresponds to a REQ signal requesting radio environment information regarding the receiving RAS, and includes its self-RAS id and null data for compatibility with an ACK signal.

 ISBN: 978-1-61208-411-4

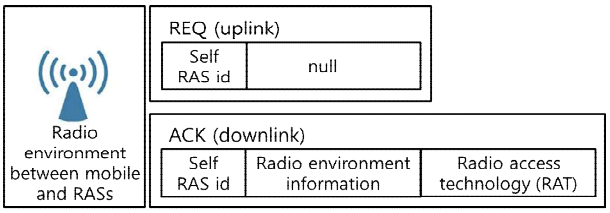


Figure 8. Message of micro-band SPC

Fig. 8 is shown as an overview of a micro-band SPC, which is similar to an in-band signal of a CPC. It refers to a signal for transmitting/receiving optimum radio environment information between a RAS and a RMS when RASs share radio environment information, which includes a REQ message and an ACK message. The REQ message refers to a message transmitted from a RMS to a RAS through an uplink, and includes a RMS id field and a null field. The ACK message refers to a message transmitted from the RAS to the RMS.

V. CONCLUSION

To expand capacity and minimize interference among macro and small cells, we proposed radio access scheme using super pilot channel in reconfigurable multi RAT-based wireless communication system, in which includes a RAS configured to share radio environment information and the multi RATs with an adjacent RAS using a macro- band SPC, the RAS being reconfigurable in conformity with the RATs. The RMS is configured to transmit and receive the radio environment information and the RATs to/from the RAS using a micro-band SPC and access the RAS using the micro-band SPC, the RMS being reconfigurable in conformity with RAT of the accessed RAS. For further study, we will set up the simulation model using our proposed radio access scheme with super pilot channel and then evaluate its result.

REFERENCES

[1] M. L. Qian, W. Hardjawana, Y. H. Li, B. Vucetic, J. L. Shi, and X. Z. Yang, “Inter-cell interference coordination through adaptive soft frequency reuse in LTE networks,” IEEE Wireless Commun. and Networking Conf., Apr. 2012.

[2] J. Hoydis, M. Kobayashi, and M. Debbah, “Green small-cell networks,” IEEE Vehicular Technology Mag., vol. 6, no.1, pp. 37~43, 2011.

[3] Z. Liu, T. Peng, H. Chen, and Wenbo Wang, “Optimal D2D User Allocation over Multi-Bands under Heterogeneous Networks,” GLOBECOM., pp. 1339 – 1344, Apr. 2012.

[4] K. H. Kim and B. W. Seo, “Efficient ICI Self-Cancellation Scheme for OFDM Systems,” ETRI Journal, vol. 36, no. 4, pp. 537-544, Aug. 2014.

[5] P. Sebastian, K. Marius, J. Martin and K. Thomas, "Ultra broadband indoor channel measurements and calibrated ray tracing propagation modeling at THz frequencies," Journal OF Communications and Networks, vol. 15, no. 6, pp. 547-558, Dec. 2013.

[6] Y. Ji, K. Zhang, Z. Feng, C. Chi, P. Zhang, “CPC-assisted Network Selection Strategy,” Mobile and Wireless Communications Summit, pp. 1-5, 2007

[7] Q. Zhang, Z. Feng and G. Zhang, “A Novel Homogeneous Mesh Grouping Scheme for Broadcast Cognitive Pilot Channel in Cognitive Wireless Networks,” IEEE International Conference on Communications (ICC), pp. 1-6, 2010

[8] J. R. Moorman, “Implementation of a 3G W-CDMA Software Radio,” ICC, vol. 4, pp. 2494-2499, 2003.

[9] R. D. Raut and K. D. Kulat, “SDR Design for Cognitive Radio,” International Conference on Modelling, Simulation and Applied Optimization, pp.1-8, 2011

[10] M. Donelli and C. Sacchi, “Implementation of a Low-Cost Reconfigurable Antenna Array for SDR-based Communication Systems,” IEEE Aerospace Conference, pp. 1-7, 2012

 ISBN: 978-1-61208-411-4

Mobile Devices Routing Using Wi-Fi Direct Technology

Ricardo Pagoto Marinho, Urbano Botrel Menegato,
Ricardo Augusto Rabelo de Oliveira

Laboratório Imobilis
Universidade Federal de Ouro Preto
Ouro Preto, MG - Brazil
Email: {ricardopmarinho,urbanobm,rrabelo}@gmail.com

***Abstract*—Information exchange between mobile devices grows every day. The communication relies on a network with an access point (Wi-Fi, cellular, etc.) using ad hoc communication because Wi-Fi Direct would avoid this dependence. Currently, Wi-Fi Direct does not support multi-hop communication or moving devices. This work focuses on expanding the use of Wi-Fi Direct technology, so that information sent from a device can walk on the network (multi-hop). We measured the number of exchanged messages by devices using four routing protocols: i) flooding, ii) Ad hoc On-demand Distance Vector (AODV), iii) AODV-Backup Route (AODV-BR), and iv) Location-Aided Routing (LAR). We see that even with some challenges one can route mobile devices over Wi-Fi Direct.**

Keywords–Wi-Fi Direct; Ad-hoc Network; Routing.

I. INTRODUCTION

Mobile devices have a highlighted place in society life. Many people have some type of mobile device. Therefore, communication using mobile devices is very typical. Apps that help them communicate are frequently used and are popular. However, these apps rely on some sort of network so that the messages can be exchanged, whether it is Wi-Fi or cellular.

These types of networks are already well known and can guarantee that information is delivered even though it may not be possible to use them. Free and good Wi-Fi networks are not easy to find or are not available [1]. In addition, the device must be on the reaching area of a Wi-Fi modem so it can connect itself to the network. Cellular networks can be found anywhere in a city in which they are available, even though access to these networks is made by means of payment whether right on time or before access.

Ad hoc networks can be a cheaper alternative for the devices to communicate. In this type of network, the devices communicate directly with no dependence on the access point as modems on Wi-Fi networks or antennae on cellular networks. In this context, Wi-Fi Alliance has developed Wi-Fi Direct technology [2]. This technology uses a Wi-Fi interface for them to communicate in an ad hoc fashion [3].

This is a new technology and few devices have it. Android devices have this technology. Because Android 4.0 devices are equipped with Wi-Fi Direct technology, we can exploit the technology to communicate between devices. Currently, the communication only happens between two devices that are within the reaching area of both with no hops on communication. Moreover, the devices must be motionless so that communication can be performed efficiently.

Botrel Menegato *et al.* [4] use Android's service API to publish information such as speed, location and battery power. This information can be shared between devices. Based on Android implementation and through experiments, we observed that there is a 24-character limit for the service name being published. We think that 24 characters in many cases are not sufficient to give a relevant name to a service. Very often, 24 characters are not sufficient so that all information on a route message can be sent. Thus, for routing algorithm experiments in some cases, information must be split into more than one publishing.

In this paper, we intend to expand this technology's use on the Android. By doing so, the mobile devices that use Android can communicate in a multi-hop and mobile network way (with devices joining and leaving). To do so, we used four ad hoc routing protocols: i) flooding, ii) AODV [5], iii) AODV-BR [6], and iv) LAR [7]. The flooding protocol is the simplest. With it, the devices only send messages to its neighbors, with any type of control. These messages allow neighbors to know that the devices are on the network. Neighbors, on their turn, replicate with any type of control the information so that every device on the network is aware of other devices. On AODV, AODV-BR and LAR algorithms, devices that want to communicate must initiate a routing discovery stage before sending any message. On this stage, the device that wants to communicate starts a route request with its neighbors. They replicate the message until it arrives on the destination device. The destination then creates the reply message. This message has the route that the request message went through to arrive at the destination. When the reply message arrives on the source device, the route is created. LAR is different from AODV because LAR uses the location and moving information of the destination device to forward requests. AODV-BR is a modification of AODV. The devices maintain a record of the reply messages from other devices so they can produce backup routes for link failure cases.

The experiments were made to measure the number of exchanged messages by every protocol using Wi-Fi Direct. During the experiments, the devices were always in the range of each other. Experiments were performed by varying the number of devices from two to seven. They were performed to determine if the technology can scale and understand its behavior as we added more devices on the network.

The contribution of this paper is to show that, even with technology limitations, we can do ad hoc routing of Android devices using Wi-Fi Direct.

This paper is organized as follows: in Section II, we show several works in literature that also perform experiments

 ISBN: 978-1-61208-411-4

on ad hoc network protocols to understand the behavior of the technologies and scenarios used. Section III shows how the protocols are proposed in the literature. In Section IV, we explain how we implemented the proposed protocols on the technology. Section V explains the configurations of the experiments. Section VI shows the results we obtained from the experiments on the technology and Section VII concludes the work.

II. Related Work

Barolli *et al.* [8] propose experiments on Mobile Ad hoc Networks (MANET) using the Optimized Link State Routing (OLSR) protocol. Using eight different scenarios for the experiments, they collect information regarding throughput, Round-trip Time (RTT) and package loss. To do so, they spent 150 seconds on each experiment in a closed environment with all devices reaching everyone. The experiments were performed using OLSR over data flow from TCP and UDP to count the metrics.

Sharma *et al.* [9] use Content Centric Networks (CCN) for communication devices on a MANET to increase the message delivery efficiency on this network. The CCN paradigm only takes into account what the information brings from where it comes. Similar to Barolli *et al.* [8], they also use OLSR to perform the experiments. The proposed algorithm relies on a Multipoint Relay (MPR). The work uses probabilities for a node that is selected based on whether it is an MPR. The experiments were performed on Android devices forming different network topologies. Information regarding the package loss rate, delivery time, network traffic and overhead was collected.

Ikeda *et al.* [10] propose experiments to evaluate throughput and package loss rate on MANETs using OLSR and B.A.T.M.A.N. protocols. The devices were used on two scenarios. One scenario has every device stationary, and other scenario has one of the devices moving. Moreover, the devices were on different floors of the building.

Won-Suk Kim and Sang-Hwa Chung [11] proposed a modification on AODV. They used multi-interface multi-channel (MIMC) wireless mesh network issues (WMN) and adapted the protocol for these situations. The new protocol was named Optimized MMIC ADOV (OM-AODV).

Oki *et al.*[12] verified how much battery power AODV and OLSR consume. They used 14 devices to perform the experiments. The goal was to verify which protocol was better for different situations on solar powered devices. The experiments showed the efficiency of these two protocols with different transmission power and information size.

Liu Yujun and Han Lincheng [13] use a modification over AODV-BR to reduce traffic load when a link failure is detected. They modify RREP messages to make an Extended Hello Message so only neighbors obtain these messages. When a link failure occurs, nodes search over their neighbors so they can find another path to the destination. This modification allows nodes to use fewer control messages to adapt to topology changes.

All of these works focus on testing and experimenting with different types of ad hoc network protocols to determine how the scenarios and technologies that use them behave when using the protocols. This is similar to the work we propose. However, our work uses a technology that to our knowledge, have never been experimented with before, namely, Wi-Fi Direct. In addition, these works differ from ours because we propose a scenario where the communication is all Peer-to-Peer (P2P), while nearly every work presented relies on a certain type of group communication.

III. Algorithms

In this section, we show how the four selected protocols are proposed. The protocols to test the technology are flooding, Ad hoc On-demand Distance Vector (AODV), AODV Backup Route (AODV-BR) and Location Aided Routing (LAR).

These protocols were selected because they are basic protocols for ad hoc network routing. Flooding is the simplest one, AODV only uses the basic route discover approach, and LAR is slightly more complex because it has positioning and movement information. AODV-BR was implemented to test our hypotheses to determine whether a modification on the protocols can make the technology work better. As discussed in Section II, most works used the OLSR protocol, and we did not implement it because these four protocols are sufficient to support our claims.

A. Flooding

For the flooding protocol, the nodes that participate in the network only reply to messages sent by its neighbors without any control on them. Thus, two non-neighbor nodes that have a common neighbor can identify themselves as belonging to the network.

However, with this algorithm, we have the guarantee that the message will go through the best path between two nodes because every node on the network will receive it.

B. AODV

The AODV protocol is a reactive protocol. On this type of protocol, a node will know if there is a route to a destination after a route discovery phase.

This phase begins when a node s wishes to communicate with another node d on the network. If node s does not have a route to d, the route discovery phase starts. First, node s sends its neighbors a Route Request (RREQ) message through a broadcast. The message has the id from s and d and the route on which the message passes through (with the source id - s). The message is replicated by the intermediate nodes through a broadcast. These nodes add their ids to the route field to form a reverse route used by the reply message to arrive on s. The reply message is created in two cases: i) the message arrives on the node d, or ii) the message arrives on a intermediate node that has a route to d.

In these two cases, a Route Reply (RREP) message is created. This message has the id form s and d and the route on which the RREQ message went through to arrive on the node and where the RREP message must return to arrive on s. When an intermediate node has a route to d, it adds its id and the route that it has to d on the message.

Figure 1 shows the first process. First, node F wishes to create a route to node D. Node F sends an RREQ message to its neighbors, in this case, node X with its id - F. Node X adds its id to the route and sends the message to D. When node D receives the message, it creates an RREP message and sends it through the reverse route. Then, it sends the message to X and then to F, creating the route.

On the second process, X already has a route to D. When it receives an RREQ message from D, it just appends its route

on the message route field and creates an RREP message to *F*.

If some node leaves the network whether by a connection problem or it really leaves the network, a Route Error (RRER) message is created. The message is sent by the node that identified the problem to notify the nodes that participate in the network so that the problematic node is not available. Thus, as soon as a node receives this message, it removes every route that has the broken node from the routing table.

C. AODV-BR

This protocol is a modification of classic AODV. With this modification, the nodes produce alternate routes to a destination. When a node receives the RREP message, it stores the route information from the message, even if it already has a route to the destination. With this, other routes to the destination are formed when a link failure occurs.

In this paper, we use this idea to obtain route information from RREP messages. However, unlike AODV-BR, we use the information for a node to decide if it will send its RREP message.

When the node receives some RREP message, it stores information regarding the node that the message is for - who has begun the route discovery, and the hop numbers for that route.

With this information in hand, when the node wishes to send an RREP message, it first checks if there is already a route for that destination.

When a route is not known, the node sends its message. However, if a route is already known to the destination, the node checks if its route is better - in number of hops - than the ones it has stored - received from other nodes. If so, it sends the message. If not, it does not send it.

D. LAR

The LAR protocol, similar to AODV, is a reactive protocol. It relies on RREQ and RREP messages to create routes even though it differs from AODV by using devices' geographical positioning information on the network to communicate.

When a device s wishes to communicate with another device d, it first needs information of where, at t_1 instant, d was. This information is geographical positioning information (latitude - X_d - and longitude - Y_d), direction (D_d) and movement speed (V_d). Based on this information, when s wishes to make a route to d on a t_2, instant it calculates a possible area where d can be on t_2. This area is a circle centered on X_d and Y_d with a radius $V_d(t_2 - t_1)$.

Once this area is created, a request is sent to s's neighbors. When one of its neighbors receives this message, it verifies whether it is on the specified area. If it is, it keeps propagating the request; if not, it discards the message. The reply happens the same way on AODV.

To make more devices propagate the request, we can increase the possible areas where d might be. This area makes a rectangle, where s and the possible area that d might be are on a diagonal. In this way, the rectangle will have (X_s,Y_s), $(X_d+V_d(t_2-t_1), Y_s)$, $(X_d+V_d(t_2-t_1),Y_d+V_d(t_2-t_1))$ and $(X_s, Y_d + V_d(t_2 - t_1))$ as the edges. These coordinates have where s is and the circle that d might be: $(X_d + V_d(t_2 - t_1)$ and $Y_d + V_d(t_2 - t_1))$.

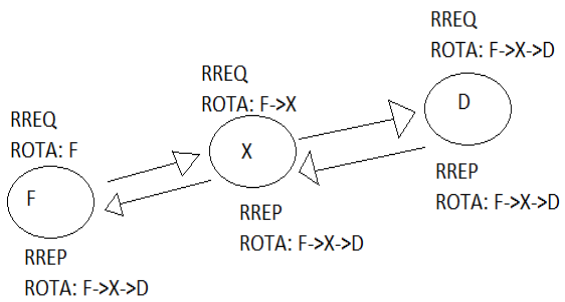


Figure 1. Route creation process.

IV. IMPLEMENTATION

To test the technology on the proposed environments, we used Android, Wi-Fi Direct, and service publishing functions.

The Wi-Fi Direct technology offers functions to recognize nearby devices using it, as search and automatic identify, and elects a Group Owner (GO) to manage the network [4].

A. Framework

Botrel Menegato *et al.* [4] created a service publishing framework to elect cluster heads. So, they wish to make the GO election more reliable for the elected node relevant inside the cluster context.

They used all functions offered by the Android API to Wi-Fi Direct and service publishing, including scanning the network for devices and receive messages sent by neighbors. Thus, the information regarding the cluster head was sent to neighbor nodes so they could decide which device was the best for the job.

Information regarding devices on the network was published for available services they had. These published details are strings with the information that devices wish to send, even though it has size limitations that make it possible to send only a small amount of information (approximately 24 characters) at once. In addition, published services are continuously sent by the API, ending only they when are explicitly requested. Other limitation is that there is a limit on the number of different services being published by one device. Each one can have approximately seven different services. Once this number is reached, it is necessary to end some service publications to start another one.

Information published in [4] was only sent to neighbors and to those that do not send information to their neighbors. This issue opens another use for the framework. We can use it where information must be sent to other nodes beyond neighbors such as in routing.

B. Protocols

To implement the protocols, we made every message they must send a service that is published by the device. This means that when a device wishes to create a route to another device, the RREQ message is a service with all information on it, such as the type of message - RREQ, the source and destination device ids and the route. When an intermediate device receives and forwards a message, such as a device forwarding an RREQ message, it also publishes the message it is making available as a service, even though the process was not initiated by it. So, when this occurs, the device has its own services, as requests and replies initiated by it, and services from another device published.

 ISBN: 978-1-61208-411-4

When choosing a device to initiate a request, the devices have a table that stores ids from other devices on the network. This table is filled when a device enters the network and publishes a service (message), informing the other devices that it is on the network. With this table containing at least one id on it, the device starts the route request process. When more than one id is on the table, it chooses at random the one to which the request will be made.

As we said before, the service publishing API available for Android has limitations. When a routing information does not fit in only one message (has more than 24 characters), it is split into additional messages so that all of the information can be sent. Putting this together with the services that continue being published by the API, the network has more load.

Thus, to build the RREQ and RREP messages, one must use numbers to show the number of messages that will be necessary to send the information and to show where the received message is, together with the usual routing information as source, destination and route. For example, if a message must be split into two, the first one has the numbers two (total amount of messages) and one (showing that this is the first one). The second has the numbers two and two (total amount and second message).

When a device receives a message with a number greater than one for the total number of messages to be received, it keeps this information until the next one arrives. As soon as it arrives, the information is stored.

A problem that may occur is the messages arriving out of order. In this case, the second message is stored and completed when the first one arrives.

In addition, because the LAR implementation has more information than AODV, such as geographical positioning and direction information, we had to split it into additional messages so all of the information could be sent, causing more challenges to the received message control.

To use the work from [4], we had to expand its functionalities. So, it was necessary to make modifications when a device receives a request and sends the message, as discussed in Section III.

V. EXPERIMENTS

Experiments were performed to verify Wi-Fi Direct's scalability when we introduce more devices for communication. For this, we used from two to seven tablets: five Samsung Galaxy Tab 2 (three of them as an operational system Android 4.1.1 and two as an Android 4.1.2) one Samsung Galaxy Tab 3 with Android 4.2.2 and one Samsung Galaxy Note with Android 4.3.3.

The experiments were performed with every tablet in the range of all of them so that the scalability could be tested. On the experiments, we counted the number of messages exchanged on each one of the protocols previously cited to test how the technology behaves with them.

Every time an experiment was performed, we switched off the Wi-Fi interface for the messages to stop publishing, avoiding the interference with the next experiment. After this, the application on the tablets was initiated and the message number was counted. Each experiment took between 15 and 20 minutes to be completed.

VI. RESULTS

Here, we show and discuss the measuring results by graphics.

A. Flooding

On the flooding protocol implementation, we only used one type of message. This message tells the neighbors about the existence of a node on the network. As previously discussed, as soon as the neighbors receive this message, they reply to their neighbors, so every node on the network can be aware of the participating nodes on the network. By doing this, the measures were made by considering the number of sent and received messages for a node. Figure 2 shows the means of the sent messages by tablets for the experiments.

By analyzing Figure 2, one can see that as we put more tablets on the network, the sent messages mean keeps growing. This shows that the technology bears the introduction of more devices on the network for sent messages.

We can also see that this message increase is nearly linear. This shows that up to where we made measurements, the technology does not modify its behavior as we introduce new devices. The figure also shows the standard deviation on the measures. They are small and tell us that the variation on the measures was small. Because the variation is small for every measurement, we can state that the technology does not change its behavior as we introduce more devices on the network. This confirms our affirmation that for this type of message, the technology bears the introduction of more devices on the network and is stronger.

By analyzing Figure 3, one can see that for up to four tablets on the network, the mean rises linearly. This shows that the technology bears, without any problem, received messages from four tablets at the same time. With more tablets, the network starts to deteriorate, *i.e.*, lose its ability to receive messages, and with six and seven tablets, the deterioration was greater.

The network deterioration can be better shown in Figure 4. As previously discussed, the technology sends many messages until the service is ended. This makes the number of received messages great than sent messages.

With this said, we made Figure 4 by using the difference between the number of received and sent messages. Along with everything that was discussed in the previous paragraph, we can wish that for five tablets, more messages are received than sent so that the figure will have a negative number. At this point, the difference between sent and received messages should rise and the numbers should become lower. However, for six or more tablets, for the point where the network deteriorates, this number should be positive.

As we thought, the difference between sent and received messages grew up to five tablets after this because the network deteriorates the difference invert and more messages are sent than received.

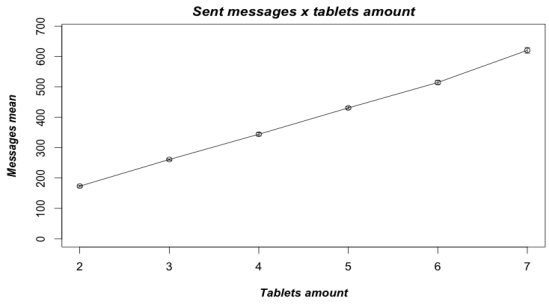
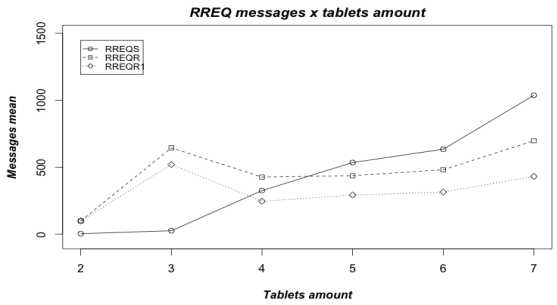
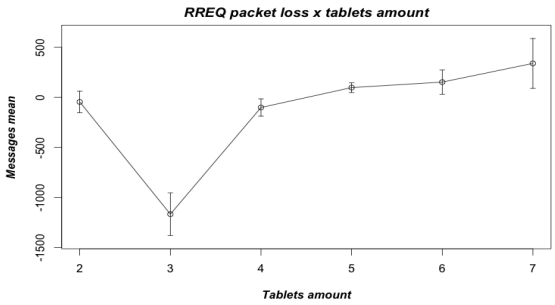


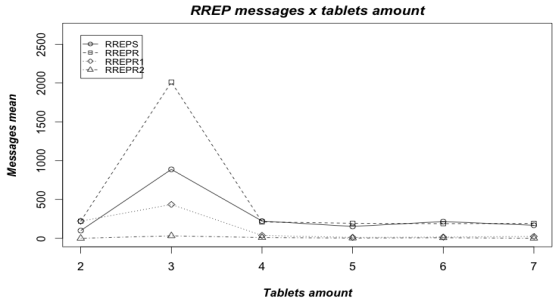
Figure 2. Sent messages x Tablets number on flooding.



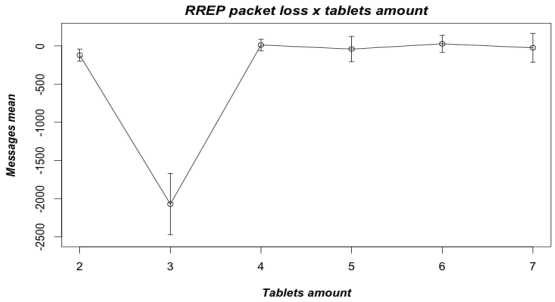
(a) Comparison between shared RREQ messages on AODV.



(b) RREQ packet loss x Tablets number on AODV.



(c) Comparison between shared RREP messages on AODV.



(d) RREP packet loss x Tablets number on AODV.

Figure 5. AODV comparisons.

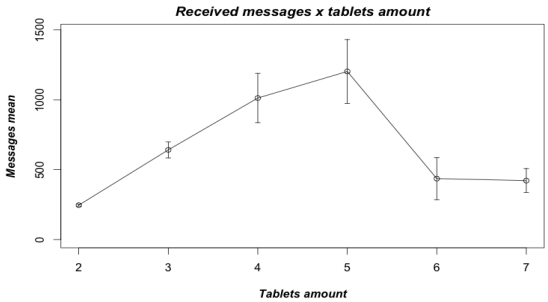


Figure 3. Received messages x Tablets number on flooding.

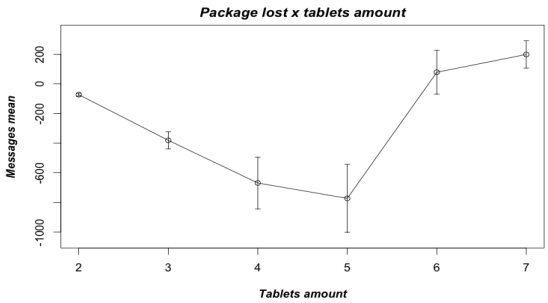


Figure 4. Packet lost on flooding.

B. AODV

AODV has three types of messages, namely, RREQ, RREP, and RERR. RERR messages were not taken into account because every tablet was in the range of all of them, and this type of message is not generated very oftem. RREQ and RREP were split into several categories because all of the messages are not applicable to a device. An example is an RREQ message arriving at a device, while the request was not for it. So, the RREQ message measures were split into sent (RREQS), received (RREQR), and received messages on which the destination was the device. This message was tagged as type 1 received RREQ(RREQR1).

Not every RREP message is relevant for a device. Consider the case where a device receives an RREP message, but the device is not on the reverse route; the message has no meaning for it. So, they were split into measurements for sent (RREPS), received (RREPR), received so that the device made the request (type 1 - RREPR1) and received so that the device did not make the request but belongs to the reverse route (type 2 - RREPR2).

Figure 5a shows a comparison for RREQ messages. For received RREQ, one can see that the network begins to deteriorate after three tablets are on it. Here, the tablets receive nearly all of the sent RREQ messages. However, after this point, all received message measures drop. An interesting fact is that with seven tablets, the network received nearly the same number of messages, when there are three tablets on it. This can show us a technology boundary. To confirm this statement, experiments with more tablets are needed.

Type 1 received RREQ messages have the same behavior as the received RREQ messages. This occurs because type 1 received RREQ messages are within the received RREQ messages.

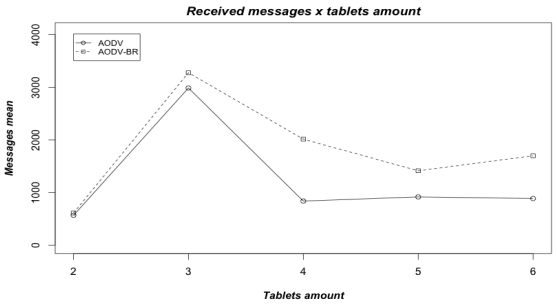
Comparing the curves in Figure 5a, when we introduce more than four tablets, the network begins to deteriorate. While the sent messages curve always grows, the curve of the received messages drops when it arrives at four tablets and starts to grow when we have more tablets on the network.

Figure 5b shows the difference between sent and received RREQ messages. We hope that for two and three tablets, the curve grows negatively and for four or more tablets, it begins to grow positively because of network deterioration.

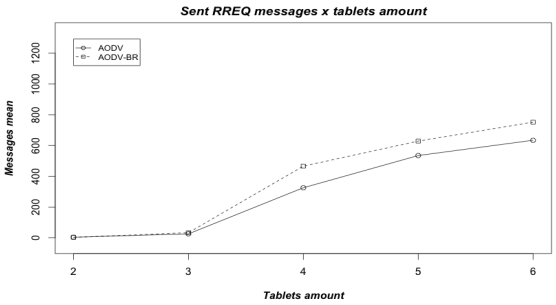
As previously discussed for two and three tablets, the technology receives the most messages and their copies are made by the service API. With five or more tablets, the number of sent messages is greater than the received messages.

Figures 5c and 5d show the same comparisons made with RREQ messages for RREP messages.

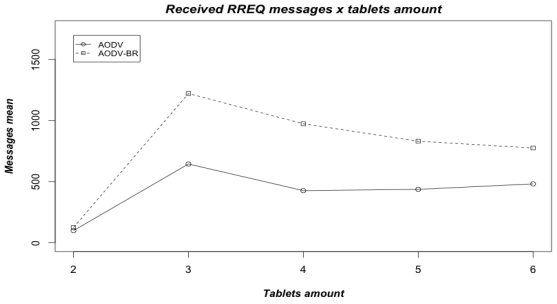
 ISBN: 978-1-61208-411-4



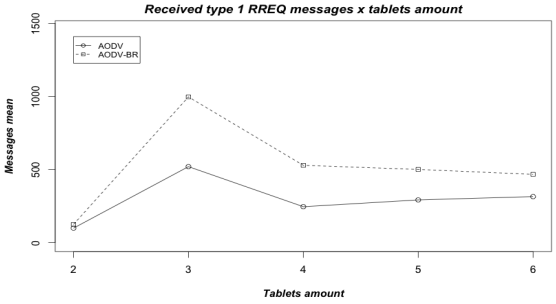
(a) Received messages number x Tablets number.



(b) Sent RREQ messages x Tablets number.

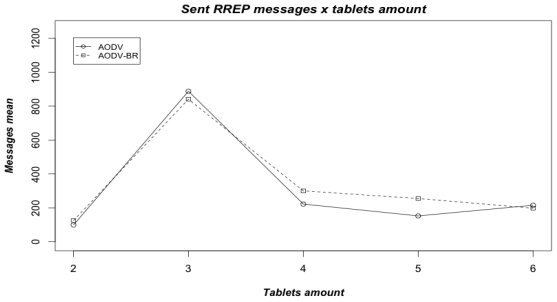


(c) Received RREQ messages x Tablets number.

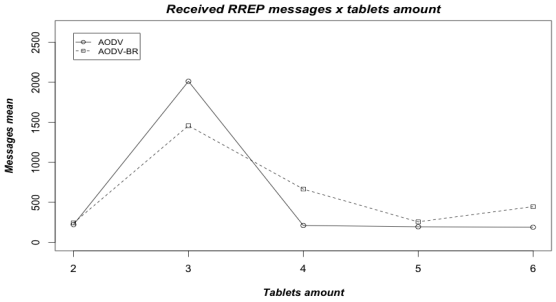


(d) Received type 1 RREQ messages x Tablets Number.

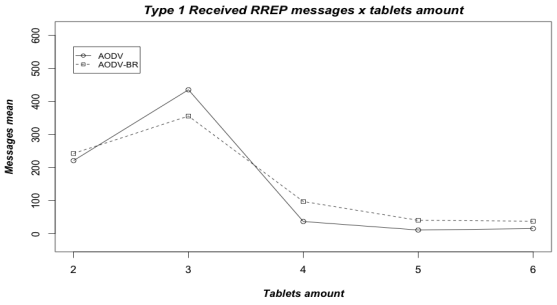
Figure 6. AODV/AODV-BR comparison - 1.



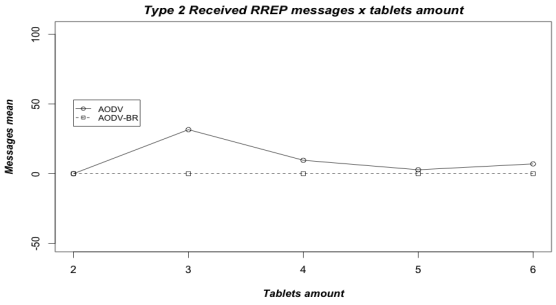
(a) Sent RREP messages x Tablets number.



(b) Received RREP messages x Tablets number.



(c) Received type 1 RREP messages x Tablets number.



(d) Received type 2 RREP messages x Tablets number.

Figure 7. AODV/AODV-BR comparisons - 2.

By analyzing Figure 5c, one can see that for every type of message, its mean drops when there are four tablets on the network. This differs from flooding and RREQ messages measures. This occurs because RREP messages are created only after an RREQ message arrives. As Figure 5a shows, when there are four tablets, the receiving process of the RREQ messages is affected, causing problems for the RREP messages. Moreover, Figure 5c shows that with four or more tablets, the sent messages are nearly equal to the received messages. We can state that with four tablets the network begins to deteriorate and affects message receiving. Figure 5d shows this comparison in a better way. This figure shows that the network receives the most messages with three tablets and that this is the lower measured value. From this point on, the network begins to deteriorate and the comparison differences are closer. This figure's behavior differs from others because the messages rely on receiving RREQ messages.

C. LAR

For LAR experiments, we realized that Wi-Fi Direct could not receive and send messages for a long period of time (it worked for approximately 1 minute).

Because the API has limitations, *e.g.*, the number of

services published, we performed experiments by putting a limit on this number to determine how long Wi-Fi Direct would work. The results are in Table I.

The table shows a possible limit on the Wi-Fi Direct working period when used for routing purposes: 15 minutes. This value was seen for the AOVD experiments, while on flooding, the time was 20 minutes. We can state that the amount of service being published at the same time affects the technology working time.

D. Comparison: AODV/AODV-BR

Here, we will show the results when we compare AODV and AODV-BR. Our goal in making this modification is to decrease the number of messages and keep the network working better for more time, *i.e.*, prevent the network from deteriorating with four tablets or make the deterioration smoother.

Figure 6a shows the difference between received messages on AODV and AODV-BR. It shows that on both protocols, the network has deteriorated when we have four tablets on it. However, with the modification on AODV-BR, one can see that the deterioration was smoother and that with six tablets, the number has grown.

Figures 6b, 6c and 6d show a comparison of sent, received and type 1 received RREQ messages. For sent RREQ message, AODV-BR increased the number of messages. This also happens with the others types of messages, but for received and type 1 received messages, the deterioration was smoother because it occurred with received messages.

Figures 7a, 7b, 7c and 7d show the measuring results with RREP messages. The results were better than the results with RREQ messages. On every figure, one can observe that the number of messages for three tablets was lower on AODV-BR than AODV. We can see that with four tablets, the deterioration was smoother. An interesting point to observe is that on Figure 7d, there were no messages on AODV-BR. With this modification, the device is not concerned about this type of message, when everyone is on everyone else's range.

VII. Conclusion and Future Work

Analyzing the results and graphics, we can conclude that for small networks (two or three devices), the technology can bear the routing load. With more devices, routing is affected because of the great load on the network. However, with some modifications on the algorithms, such as making the devices aware of other routes in AODV-BR, this issue can be solved. However, even with these problems, it is possible to route devices on an ad hoc network using Wi-Fi Direct technology.

In addition, we found that the technology is affected when we publish different services at the same time by making it stop publishing and receiving information.

A technological contribution is suggested in that the number of characters on the name of the service being published should be greater than 24.

For future work, we will experiment whit the technology in new scenarios, such as for devices moving, joining and leaving the network and for protocols that take other parameters into consideration, including social parameters.

Table I. Amount of service x working period on LAR.

Quantidade de serviços	Tempo (min)
Sem limitação	1
10	5
8	15
6	15

References

[1] L. V. Hoang and H. Ogawa, "A platform for building ad hoc social networks based on wi-fi direct," in Consumer Electronics (GCCE), 2014 IEEE 3rd Global Conference on, Oct 2014, pp. 626–629.

[2] "Wi-fi direct | wi-fi alliance," http://www.wi-fi.org/discover-wi-fi/wi-fi-direct, retrieved: January, 2014.

[3] "Wifi p2p technical specification v1.2," http://pt.scribd.com/doc/215283500/WiFi-P2P-Technical-Specification-v1-2, retrieved: January, 2015.

[4] U. Botrel Menegato, L. Souza Cimino, S. E. Delabrida Silva, F. A. Medeiros Silva, J. Castro Lima, and R. A. R. Oliveira, "Dynamic clustering in wifi direct technology," in Proceedings of the 12th ACM International Symposium on Mobility Management and Wireless Access, ser. MobiWac '14. Montreal, QC, Canada: ACM, 2014, pp. 25–29.

[5] C. Perkins and E. Royer, "Ad-hoc on-demand distance vector routing," in Mobile Computing Systems and Applications, 1999. Proceedings. WMCSA '99. Second IEEE Workshop on. IEEE, Feb 1999, pp. 90–100.

[6] S.-J. Lee and M. Gerla, "Aodv-br: backup routing in ad hoc networks," in Wireless Communications and Networking Confernce, 2000. WCNC. 2000 IEEE, vol. 3. Chicago, IL: IEEE, Sep 2000, pp. 1311–1316.

[7] Y.-B. Ko and N. H. Vaidya, "Location-aided routing (lar) in mobile ad hoc networks," Wirel. Netw., vol. 6, no. 4, Jul. 2000, pp. 307–321.

[8] L. Barolli, M. Ikeda, F. Xhafa, and A. Duresi, "A testbed for manets: Implementation, experiences and learned lessons," Systems Journal, IEEE, vol. 4, no. 2, June 2010, pp. 243–252.

[9] P. Sharma, D. Souza, E. Fiore, J. Gottschalk, and D. Marquis, "A case for manet-aware content centric networking of smartphones," in World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2012 IEEE International Symposium on a, June 2012, pp. 1–6.

[10] M. Ikeda, E. Kulla, M. Hiyama, L. Barolli, and M. Takizawa, "Experimental results of a manet testbed in indoor stairs environment," in Advanced Information Networking and Applications (AINA), 2011 IEEE International Conference on, March 2011, pp. 779–786.

[11] W.-S. Kim and S.-H. Chung, "Design of optimized aodv routing protocol for multi-interface multi-channel wireless mesh networks," in Advanced Information Networking and Applications (AINA), 2013 IEEE 27th International Conference on, March 2013, pp. 325–332.

[12] O. Oki, P. Mudali, M. Mutanga, and M. Adigun, "A testbed evaluation of energy-efficiency of routing protocols in battery-powered wireless mesh networks," in AFRICON, 2013, Sept 2013, pp. 1–7.

[13] L. Yujun and H. Lincheng, "The research on an aodv-brl to increase reliability and reduce routing overhead in manet," in Computer Application and System Modeling (ICCASM), 2010 International Conference on, vol. 12. IEEE, Oct 2010.

 ISBN: 978-1-61208-411-4

Integrating an Effective VoIP Service in a USRP/GNU Radio Testbed

Naceur Malouch

Université Pierre et Marie Curie - UPMC Sorbonne Universités
Laboratoire d'Informatique de Paris 6 LIP6/CNRS
Paris, FRANCE
Email: Naceur.Malouch@lip6.fr

***Abstract*—This work presents an implementation of spectrum handoff and selection in order to experiment a Voice over IP (VoIP) application over a cognitive radio testbed. VoIP connections use both WiFi bands and GSM bands to select available channels dynamically. We use the USRP/Gnu radio testbed to implement spectrum management functions and to run the experiments. We also experiment VoIP and data transfers together by implementing a satisfaction-based inter-service cohabitation strategy. Through subjective observations and objective measurements, we found that the quality of established phone calls can be maintained. Besides, VoIP can benefit from the flexibility of data transfers to perform handoffs more adequately. In the same time, data transfer is able to sustain its required average rate with minimum degradations.**

Keywords–Cognitive Radio Networks; VoIP; USRP; GNU Radio; Testbed.

I. INTRODUCTION

Providing classic services such as Voice over IP (VoIP) over Cognitive Radio Networks (CRNs) has become an interesting research topic during past few years. The first challenge is to guarantee uninterrupted services despite the dynamic nature of the spectrum. Unlike traditional Quality-of-service (QoS) mechanisms for wired networks which mainly depend on the traffic statistics, the quality of required services on CRN must be according to the spectrum properties as well. The second challenge is to evaluate the suitability of a new available spectrum for usage and selecting the best channel among multiple available channels. This challenge concerns channel selection and handoff algorithms.

In this paper, we focus on supporting a VoIP application [1] over a GNU Radio testbed [2] that uses Universal Software Radio Peripheral (USRP) devices [3]. Indeed, several works have studied theoretically and by simulations the feasibility of supporting such service [4]–[9]. Other works have evaluated the general performance using experimental testbeds without considering VoIP or real-time applications [10]–[12]. [12] studies latency between USRP devices and determines reasons of large latencies at the PHY layer. To the best of our knowledge, the most relevant work that has focused on experimenting VoIP over a cognitive radio testbed is [13]. However, the main objective of this work is to study the impact of spectrum sensing on the quality of VoIP. In our work, we rather study the impact of spectrum handoff and spectrum selection which is complementary to the previous work. The previous work has limited the tests to be only over 5 GHz frequency band. We use in this work both GSM and Wi-Fi bands using several channels. In particular, we show that VoIP works also when using low frequency bands. In addition, we use a real-world VoIP application [1] rather than a traffic generator as in [13]. The testbed also is different since [13] have performed their tests using the WARP platform [10]. Finally, in this work, we also investigate the cohabitation between VoIP and data transfer in order to share the spectrum while satisfying simultaneously both service requirements.

The lack of VoIP experiments over GNU Radio testbeds is mainly due to the difficulty of running a complete TCP/IP stack over existing testbeds. Besides, once simple frame transmissions are successfully set, usually no further investigations are done for upper layers.

During the experiments over the USRP/Gnu radio testbed, we evaluated the audio quality and measured the delay and the jitter for VoIP. The results of the tests show that it is possible to compensate the delay increased by the interruption of the transmission due to the presence of a Primary User (PU) on the channel, by performing spectrum handoff when it becomes a must. Here, finding adequate metrics are necessary to find when the handoff must be performed and which target channel to select for the handoff. These metrics are related on one hand to the service requirement, and on the other hand to the channel properties such as the availability ratio and the remaining availability period. We also experimented the efficiency of using prediction techniques to assess available periods.

The rest of the paper is organized as follows. In Section II, we introduce the testbed used in the experiments and we present related settings and constraints. In Section III, we show results of the experiments over the testbed for VoIP with and without the presence of data traffic. In Section IV, we conclude the paper and point out future directions.

II. TESTBED SETTINGS AND CONSTRAINTS

GNU Radio is an open source tool-kit that provides all functionalities to handle the radio interface and process radio signals at the software level. Gnuradio transceivers [2] are composed of many elements similar to hardware domain, like filters, demodulators, decoders, etc., called blocks.

The USRP device [3] is a radio hardware that GnuRadio can tie with. We have used USRP1 which is one of the numerous versions of USRP. The USRP1 platform can support two daughterboards The GnuRadio-USRP testbed that our experiments will be run on is shown in Figure 7.

 ISBN: 978-1-61208-411-4

TABLE I. LIST OF FREQUENCIES AND THEIR AVAILABILITY UNAVAILABILITY PERIODS

Frequency (MHz)	Avail. range (sec)	Unavail. range (sec)	Avail. percentage
905	2-8	42-48	10%
915	9-15	25-31	30%
925	3-9	10-18	30%
2485	27-33	23-37	50%
2490	3-17	5-15	50%
2495	11-17	4-8	70%
2900	20-36	6-26	70%

The platform is composed of two USRP1 with two daughterboards each, RFX900 for GSM band, and RFX2400 for WiFi band and two Linux host machines, with GnuRadio installed on each one, connected via USB2.0 interface to the USRPs.

We have tested several supported modulation methods such as PSK, QPSK, GFSK, QAM, BPSK, and GMSK for both GSM and WiFi. Through the tests, gmsk gave the best results in terms of packets delivery and packet error rate.

GnuRadio and USRP do not allow straightforwardly to run simultaneous transmissions. However, this feature is necessary so that we can run more than one service simultaneously either on the same channels or on different ones. But USRP is designed to support at most two channels in one direction, one on each daughterboard.

In our work, the sensing module is the source of the information that we have to consider to decide how to manage spectrum mobility. The experiments are based on generating the sensing information using different random patterns corresponding to the availability and the unavailability of the channels. The benefit of this method is that we can compare results by using the same sets of availability patterns. Using realistic primary transmissions is not controllable and thus it is hard to compare selection and handoff algorithms fairly.

To generate the available/unavailable periods, we use the uniform distribution to control their durations and thus primary activity. For available/unavailable periods generation, we have set intervals, means and ranges of time to satisfy given percentages of availability for each channel. (Table I).

We implemented a realistic sensing-transmission cycle where the transmission operation will alternate with the sensing process periodically. We also implemented a GNU radio module in which we can plug any spectrum and handoff algorithm.

III. VOIP SPECTRUM SELECTION AND HANDOFF EXPERIMENTS

A. Experiments with VoIP alone

In order to generate different VoIP patterns, we establish real VoIP communications using mumble software while varying the duration of talkspurts and silence periods. Four types of experimentation are made during 10 minutes. The first type is with continuous communication without silence periods (Nosilence). The second one imitates a dynamic conversation. It is done with 10s for talkspurts and 10s for silences (10-10). The third represents also symmetric traffic but it is less dynamic with 60s for talkspurts and 60s for silences (60-60). The last one is with 20s for talkspurts and 120s for silences (20-120) to imitate a conversation where a speaker is listening more than speaking.

VoIP connection should select the channel with the greatest *predicted* remaining available period among the channels that have a residual bandwidth (capacity) larger than the codec rate. Besides, in order to evaluate the impact of using predicted values on the performance, we also perform experiments using the exact future remaining availability durations. This can be done since the channel patterns are generated in advance. Also, this can be useful for real systems where availability periods are known in advance through for instance regional databases. In case of a handoff, we select also the channel with the greatest predicted remaining availability period. We have used *Autoregressive Model Based Prediction (ARM)* [14] in our implementation.

First, Figures 1 and 2 show the variation of frequencies selected with spectrum management algorithms for VoIP. Since we assign more availability to WiFi channels (Table I), we notice that the algorithms tend to choose these channels more often. This is because the remaining available period of these channels is usually larger than the others. However, depending on the instantaneous availability of channels even low availability ratio channels can be used when necessary.

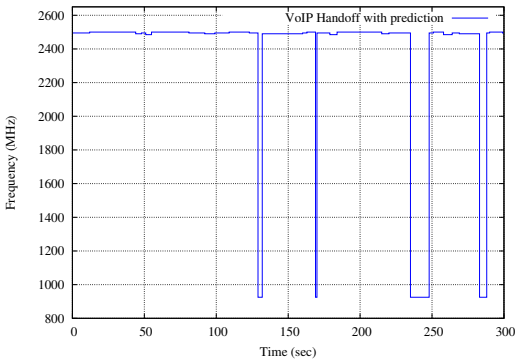


Figure 1. Channel changes during VoIP experiments

Then, we explore the number of handoffs. The results in Figure 3 are in accordance with the idea that the more

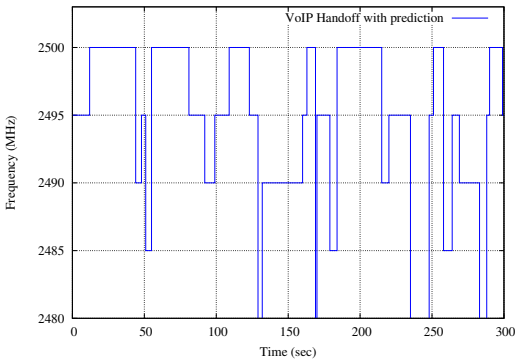


Figure 2. Zoom on channel changes over the WiFi band

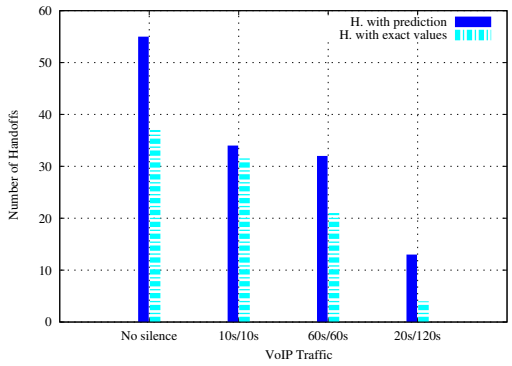


Figure 3. Number of handoffs function of VoIP traffic

important the traffic, the larger the number of performed handoffs. We notice that the algorithm with prediction gives different results than the one with exact values. This can be explained by the fact that the prediction mechanism gives sometimes inaccurate results that may lead to errors. We observe also that the prediction gives a result close to the exact values when the traffic is very dynamic because it is possible in this case to finish a talkspurt using the same channel before the arrival of a primary user. Knowing the exact value of the remaining available period is not really useful.

Notice that handoffs can occur during talkspurts. Especially, when there is no silence periods, the handoff does not impact the quality of the received audio except if the handoff is delayed. This happens when the signaling packet is lost, no channel is available immediately, or the transmission queue contains some packets. We conclude that it is better to send signaling packets using a different queue that has the scheduling priority. Handoffs however impacts a little the quality of the conversation interactivity. Globally, all tests being done, the quality can be evaluated by a subjective mean opinion score (MOS) of 3.8 since the phone calls pursue normally except few intermittent discomforts in terms of interactivity.

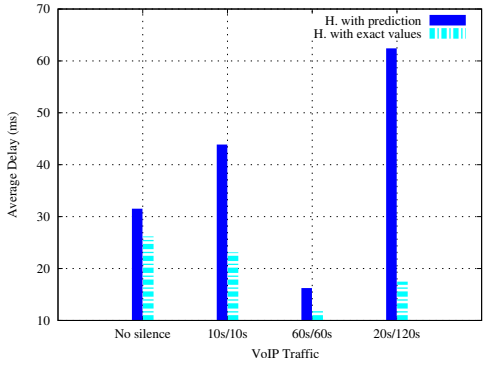


Figure 4. Average of VoIP packets delay function of VoIP traffic

From a delay point of view, we notice in Figure 4 that the algorithm with prediction has the same behavior as the one with exact values when the traffic is important (Nosilence and 60s-60s). The large delay variation noticed for some tests is caused by the fact that sometimes there are no available channels, in this case we stop the transmission (packets are buffered) until we get an available channel, and this impacts directly the delay.

The quality of the received audio confirm that if the spectrum handoff is performed immediately without any extra delay, then it does not affect packet transmissions. Indeed, we did not notice interruption of the communication. This is because the distance between the two USRP boxes is not long, hence audio packets can be delayed a little bit without a large impact. These observations confirm the idea that performing spectrum handoffs in advance when the conditions are optimal is an interesting approach to maintain a good quality of the VoIP communication. Of course, this can increase the number of handoffs and thus the energy consumption and the channel contention.

For the test 20s-120s we notice that the average of the delay exceeds 60 milliseconds. As we explained earlier, the channel unavailability causes this large delay. We draw the CDF of the packets delays shown in Figure 5. The figure zoomed on the interval [0-30]ms shows that despite that the average delay is large, most of the packets have delays less than 10ms.

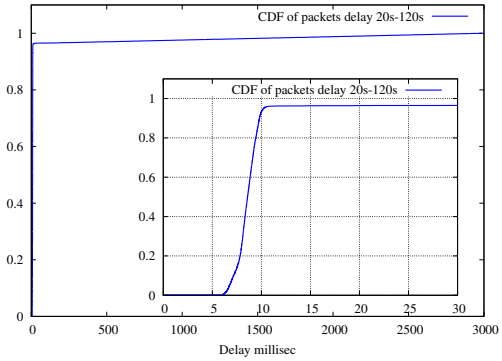


Figure 5. CDF of packets delay for 20s-120s talkspurt-silence periods and handoff with prediction

In Figure 6, when we use the exact values instead of the predicted ones for the remaining availability periods, we notice that the jitter is smaller especially when the traffic is more important because first packets of talkspurts require usually more transmission delays. Again, the algorithm with prediction has a closer result to exact values when the traffic is very dynamic and talkspurt durations are equal to silence durations (10s-10s).

B. VoIP and data transfer cohabitation experiments

In this part, each service (VoIP or Data) has to optimize its connection and guarantee at the same time the non disturbance of the other service. The Gnu Radio testbed imposes a strong constraint for the experiments: Two connections can not be established in parallel over two different channels of the same wireless card. Thus, we need to adapt the design to this constraint. The objective is to experiment the inter-service approach for cohabitation.

The inter-service approach is based on the fact that services have to consider each other. In other words, when performing spectrum selection and handoff, one service takes its decision

 ISBN: 978-1-61208-411-4

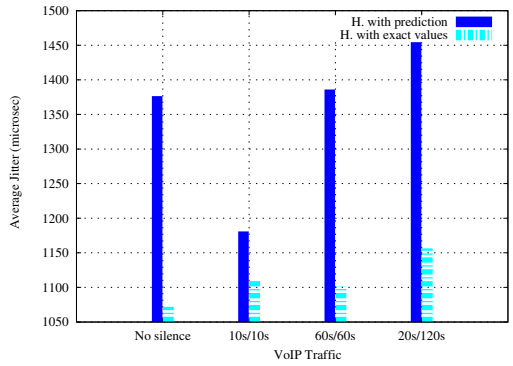


Figure 6. Average of the jitter function of VoIP traffic

not only based on its requirements but also based on demands of the other services, so that we satisfy both requirements as fairly as possible. The question here is how to assign channels to services, and how to know that this assignment is the best way to satisfy each service requirement. The basic idea is to find metrics that may describe better the state of channels for each service.

For VoIP, the goal is to find a channel that is available and allows the service to stay as long as possible on the same channel without handoff or perform the handoff in good conditions. To this aim, we can consider the remaining available period which describes only the channel instantly and does not give the global state of the channel. It is also necessary to measure the degree of satisfaction of the user. One possible metric is the mean number of handoffs that can be calculated for each channel in a given window time. Besides, this number provides somewhat a long term estimation of the channel status. However, it is difficult to relate this number to the quality perceived by the users, and also choosing the right measurement window size is not obvious.

Another possibility is to use in addition to the remaining availability period a threshold that represents a large acceptable value for the availability period. This threshold can correspond to the average VoIP duration or average duration of talkspurts. A channel that has this value of threshold for its remaining availability period is then considered as a best channel. In this case the satisfaction degree is equal to 1. In order to assess the satisfaction degree for a group of channels, we can use the maximum remaining availability period among all channels. More precisely, the satisfaction degree can be computed as follows

$$x_{VoIP_{DB_i}} = \frac{Max\ remaining\ available\ period\ on\ DB_i}{Threshold\ of\ remaining\ available\ period}, \quad (1)$$

where DB_i refers to the channels accessible through the GSM interface or the WiFi interface, $i \in \{GSM, WiFi\}$. This metric indicates whether it is worthy to stay on the current interface or it is better to move to the other interface. For instance, if at some time during the VoIP communication, the maximum remaining available period among all channels is low compared to the threshold, then in case of handoff, it is better to move to other interface to avoid more possible interruptions of the communication in the future.

For the data service, we have different parameters that we can take into account such as the available bandwidth during the availability period and the average achievable rate over the channel. The last parameter describes well the status of the channel but has again the drawback of choosing the right window size for past measurement. Alone, it is not sufficient to measure the satisfaction of the user. It should be compared to the rate demand. We use the following metric

$$x_{Data_{DB_i}} = \frac{Max\ (\frac{Available\ achievable\ rate}{Rate\ demand})\ on\ DB_i}{Tolerance} \quad (2)$$

Here, we also use the maximum value among all channels of a given wireless card. The $Tolerance$ parameter should be equal or close to 1. The smaller this parameter, the larger the tolerance to a rate decrease.

Since we have technical constraints, we have to select one channel on each wireless card and each channel is reserved for a service VoIP or data (Figure 7). Thus, few service configurations are possible and the goal is now to find what configuration to choose and when. In this case, we can calculate the fairness index for each configuration as follows:

$$\mathcal{FI}(x_1, x_2) = \frac{(\sum_{i=1}^{2} x_i)^2}{2 \cdot \sum_{i=1}^{2} {x_i}^2} \quad (3)$$

where $(x_1, x_2) = (x_{VoIP_{DB_1}}, x_{Data_{DB_2}})$ or $(x_1, x_2) = (x_{VoIP_{DB_2}}, x_{Data_{DB_1}})$. Then, we choose the configuration that has the best fairness index. In other words, we choose the configuration that tries to satisfy both services. Algorithms 1 and 2 provides the implemented spectrum handoff and selection procedures for each service.

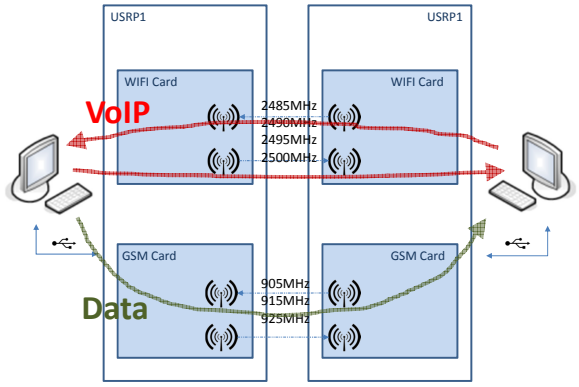


Figure 7. Example of daughterboards (wireless cards) configuration for VoIP and Data services cohabitation

Data Handoff and Selection with Cohabitation: When the data is transmitted on channel ch_c and a PU is detected, the algorithm tests if the available rate satisfies the required rate (line 3) to check if it is useful to wait. On the contrary case, we have to perform handoff immediately otherwise we wait for $t_1 = T_{OFF} + \sigma_{OFF}$ which is the sum of the average unavailability time of the current channel ch_c and the standard deviation. Then, we retest if the PU is still on the channel. This

 ISBN: 978-1-61208-411-4

Algorithm 1 Data Handoff and Selection with Cohabitation

Require: DB_c: The current daughterboard, Ch_c: The current channel, Ch_i: Channel i, AC: List of available channels on DB_c, OC: List of available channels on the other DB, T_{OFF}: The unavailability time mean of Ch_c, σ_{OFF}: The standard deviation of unavailability periods of Ch_c, Rav_i: Remaining availability time in channel Ch_i, Ar_i: Available rate of channel Ch_i, Rr: Required rate, W_{th}: Threshold weight, W_i: Weight on channel ch_i, W_{max}: Weight max on DB_c.

Ensure: Ch_{Data}: Selected channel for data, Ch_{VoIP}: Selected channel for VoIP.

1: **if** P *on* Ch_c **then**
2: $stop\ transmission$
3: **if** $Ar_c >= Rr$ **then**
4: $t_1 = T_{OFF} + \sigma_{OFF}$
5: $wait\ t_1$
6: **end if**
7: **if** $(After\ t_1\ P\ still\ on\ Ch_c)\ Or\ (Ar_c < Rr)$ **then**
8: **for all** $Ch_i\ in\ AC$ **do**
9: $W_i = Ar_i/Rr$
10: **end for**
11: $W_{max} \leftarrow max_i\{W_i\}$
12: **if** $W_{max} >= W_{th}$ **then**
13: $Ch_{Data} \leftarrow argmax_{Ch_i \epsilon AC}\{Rav_i\}$
14: **return** Ch_{Data}
15: **else**
16: $calculate\ fairness\ indexes$: $\mathcal{FI}_{Data\ on\ DB_c}$ *and* $\mathcal{FI}_{Data\ not\ on\ DB_c}$
17: **if** $\mathcal{FI}_{Data\ on\ DB_c} >= \mathcal{FI}_{Data\ not\ on\ DB_c}$ **then**
18: $Ch_{Data} \leftarrow argmax_{Ch_i \epsilon AC}\{Rav_i\}$
19: $Ch_{VoIP} \leftarrow argmax_{Ch_i \epsilon OC}\{Rav_i\}$
20: **else**
21: $Ch_{Data} \leftarrow argmax_{Ch_i \epsilon OC}\{Rav_i\}$
22: $Ch_{VoIP} \leftarrow argmax_{Ch_i \epsilon AC}\{Rav_i\}$
23: **end if**
24: **return** Ch_{Data}, Ch_{VoIP}
25: **end if**
26: **else**
27: $Go\ back\ to\ transmission$
28: **end if**
29: **end if**

strategy aims at reducing the number of handoffs performed by data transfer so that it reduces also handoffs to channels required by VoIP. Indeed, if the current channel becomes available again, then there is no need for handoff and the rate of the data transfer can still be achieved because it depends on the *average* availability. Interrupting the transfer is tolerated and does not trigger necessarily handoffs. Now, if the PU is still on the channel after t_1, we have to perform handoff. In the two cases, the mechanism of selection of the new channel is the same. We look for channels on the same daughterboard, if there is one that verifies a satisfaction metric larger than the tolerance threshold, which means that we judge that the channel may satisfy the service requirement, we move to this channel. Otherwise, we calculate the fairness index so that if we have to switch daughterboards, we guarantee the best assignment of channels to satisfy requirements of both services.

VoIP Handoff and Selection with Cohabitation: For VoIP service, the presence of PU on the Ch_c triggers an immediate handoff. The new channel is first selected on the current daughterboard if the maximum remaining availability period is greater than the threshold (lines 5-6). If not, we calculate the fairness index (line 10) to find the best configuration. In one hand, we profit from the elasticity of the data service to handoff VoIP connections whenever required. On the other hand, we do not starve totally data transmission since we use the fairness index to provide always some resources to it. This means, we can accept some degradation in the VoIP quality in order to avoid stalling totally the data transmission.

Algorithm 2 VoIP Handoff and Selection with Cohabitation

Require: DB_c: The current daugtherboard, Ch_c: The current channel, Ch_i: Channel i, AC: List of available channels on DB_c, Rav_i: Remaining availability time for channel Ch_i, Rav_{max}: Max remaining available time on DB_c, Rav_{th}: Threshold of remaining available time.

Ensure: Ch_{Data}: Selected channel for data, Ch_{VoIP}: Selected channel for VoIP.

1: **if** P *on* Ch_c **then**
2: **for all** Ch_i in AC **do**
3: $calculateRav_i$
4: **end for**
5: $Rav_{max} \leftarrow max_i\{Rav_i\}$
6: **if** $Rav_{max} >= Rav_{th}$ **then**
7: $Ch_{VoIP} \leftarrow argmax_{Ch_i \epsilon AC}\{Rav_i\}$
8: **return** Ch_{VoIP}
9: **else**
10: $calculate\ fairness\ indexes$: $\mathcal{FI}_{Data\ on\ DB_c}$ *and* $\mathcal{FI}_{VoIP\ not\ on\ DB_c}$
11: **if** $\mathcal{FI}_{VoIP\ on\ DB_c} >= \mathcal{FI}_{VoIP\ not\ on\ DB_c}$ **then**
12: $Ch_{VoIP} \leftarrow argmax_{Ch_i \epsilon AC}\{Rav_i\}$
13: $Ch_{Data} \leftarrow argmax_{Ch_i \epsilon OC}\{Rav_i\}$
14: **else**
15: $Ch_{Data} \leftarrow argmax_{Ch_i \epsilon AC}\{Rav_i\}$
16: $Ch_{VoIP} \leftarrow argmax_{Ch_i \epsilon OC}\{Rav_i\}$
17: **end if**
18: **return** Ch_{Data}, Ch_{VoIP}
19: **end if**
20: **end if**

For the tests we choose to have medium traffic for VoIP (60s-60s talkspurt-silence), two rates for the rate demand of data traffic, 100kbps and 200kbps. We do not show results for higher data rates because the VoIP software client and server (mumble) need to exchange continuously signaling packets that permit the client to stay connected. In the tests, since the number of available channels is limited, we have to wait some time for frequencies to be available, so we stop transmissions. In this case, when the client does not receive server's packets, it considers that it is no longer connected which impacts our tests especially when we increase the rate demand. The results are summarized in Table II.

The first observation is that the effective rate measured for data transfer is slightly lower than the rate demand. This is the cost to pay to keep an acceptable quality for the VoIP conversation. Indeed, the audio was not distorted during the whole duration of the communication. Besides, a lower rate

 ISBN: 978-1-61208-411-4

TABLE II. PERFORMANCE RESULTS FOR VOIP AND DATA TRANSFER COHABITATION

Required data rate (kbps)	Number of VoIP handoffs	Number of data handoffs	Effective Data rate (kbps)
100	17	14	97.9
200	21	7	180

for data transfer implies usually a little longer delay before delivering the data (file), whereas for VoIP it is crucial to maintain interactivity and audio quality in all periods of the communication.

We observe also that the number of handoffs for data decreases when we increase the rate demand. This is because less channels can provide a better achievable rate for data and since VoIP has the priority when it needs handoffs, the data algorithm decides more often to stay on the same channel. This explains also the lower measured effective data rate compared to the case where the rate demand is lower. On the other hand, VoIP continues profiting from handoffs to choose the adequate wireless card and channel to pursue its communication. This confirms again that exploiting data flexibility is a practical approach to cohabit VoIP and data together.

IV. CONCLUSION

Cognitive Radio presents the perfect solution for many spectrum scarcity problems in several areas as long as classic services can be supported. Spectrum selection and handoff can achieve the quality required by these services when they experience degradations because of the appearance of primary users on the ongoing channel.

We have experimented real-world VoIP communications over the USRP/Gnu radio testbed in which we have implemented suitable spectrum selection and handoff algorithms. Our first observation is that even low availability ratio channels are useful to maintain VoIP calls through spectrum handoff without impacting substantially the quality. Moving from GSM band to Wi-Fi band and inversely can be done during the VoIP call. The spectrum handoff is performed to the channel with largest predicted remaining available period. We found that, the more the conversation is active (small talkspurts, small silence periods), the less the impact of prediction errors. However, when the talkspurts are large, it is better to perform the handoff in advance before the primary arrival to avoid abrupt interruptions.

We have also proposed a strategy and algorithms for cohabitation between two different services VoIP and data transfer. The results of experiments show that VoIP can benefit from the elasticity of data transfers in order to perform handoffs and choose adequate channels more easily. To do so, data transfer should avoid systematic handoff at primary arrival and do handoff to free the ongoing channel if required by a VoIP communication, while keeping an acceptable quality for itself in terms of achieved average rate. It is important to notice that a simple priority mechanism for VoIP as in traditional wired or wireless communications is not suitable to take spectrum decisions. Our mechanism is rather based on compromising satisfaction degrees of both VoIP and data transfer services.

More generally, the results of this work have demonstrated that acceptable quality for stringent services can be ensured in presence of primary users with dynamic activity. To go further, it is interesting to experiment the cohabitation with other types of services such as video streaming and remote desktop. In this case, the satisfaction degree and the spectrum handoff strategy of VoIP should not change. Also, the next step is the deployment of this testbed for everyday phone calls in a local network so that more statistics and more subjective opinion scores can be collected.

REFERENCES

[1] "Mumble: Audio and VoIP Application (LightSpeed Gaming LLC Company)," 2014, URL: http://www.mumble.com [accessed: 2015-05-12].

[2] "GNURadio: The Free and Open Software Radio EcoSystem," 2014, URL: http://gnuradio.org/redmine/projects/gnuradio/wiki [accessed: 2015-05-12].

[3] "ETTUS Research, USRP: Universal Software Radio Peripheral," 2014, URL: https://www.ettus.com/product [accessed: 2015-05-12].

[4] S. Lirio Castellanos-Lopez, F. Cruz-Perez, M. Rivero-Angeles, and G. Hernandez-Valdez, "Joint connection level and packet level analysis of cognitive radio networks with voip traffic," IEEE Journal on Selected Areas in Communications, vol. 32, no. 3, March 2014, pp. 601–614.

[5] S. Castellanos-Lopez, F. A. Cruz-Perez, M. E. Rivero-Angeles, and G. Hernandez-Valdez, "Performance comparison of voip cognitive radio networks under on/off and poisson primary arrivals," in IEEE 24th International Symposium on Personal Indoor and Mobile Radio Comm. (PIMRC), Sept 2013, pp. 3302–3307.

[6] Z. Wang, T. Jiang, L. Jiang, and X. He, "Voip capacity analysis in cognitive radio system with single/multiple channels," in 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM), Sept 2010, pp. 1–4.

[7] H. Lee and D.-H. Cho, "Voip capacity analysis in cognitive radio system," IEEE Communications Letters, vol. 13, no. 6, June 2009, pp. 393–395.

[8] T. Chakraborty, I. Misra, and S. Sanyal, "Selection of optimal transmission time in cognitive radio network for efficient voip performance," in 5th International Conference on Computers and Devices for Communication (CODEC), Dec 2012, pp. 1–4.

[9] H. Lee and D.-H. Cho, "Capacity improvement and analysis of voip service in a cognitive radio system," IEEE Transactions on Vehicular Technology, vol. 59, no. 4, May 2010, pp. 1646–1651.

[10] "Rice University WARP: Wireless Open-Access Research Platform," 2015, URL: http://warp.rice.edu/ [accessed: 2015-05-12].

[11] K. T. et al., "SORA: high-performance software radio using general-purpose multi-core processors," USENIX Symposium on Networked Systems Design and Implementation (NSDI), April 2009, pp. 75–90.

[12] N. Truong, Y.-J. Suh, and C. Yu, "Latency analysis in gnu radio/usrp-based software radio platforms," in IEEE Military Communications Conference, MILCOM 2013, Nov 2013, pp. 305–310.

[13] K. Tan, K. Kim, Y. Xin, S. Rangarajan, and P. Mohapatra, "Recog: A sensing-based cognitive radio system with real-time application support," Selected Areas in Communications, IEEE Journal on, vol. 31, no. 11, November 2013, pp. 2504–2516.

[14] Z. Wen, T. Luo, W. Xiang, S. Majhi, and Y. Ma, "Autoregressive spectrum hole prediction model for cognitive radio systems," in IEEE International Conference on Communications Workshops, May 2008, pp. 154–157.

ACKNOWLEDGMENT

Many thanks to Nada Abdelkader for all her works over the testbed and especially for solving parallel transmission issues. We also thank Jawad Seddar from Thales Communications and Martin Peres from LaBRI for their valuable remarks regarding the testbed settings. This work is partially supported by the French national research agency (ANR), LICORNE Project, grant ANR-10-VERS-005-03.

 ISBN: 978-1-61208-411-4

Cell Deployment Optimization for Cloud Radio Access Networks using Teletraffic Theory

Andrijana Popovska Avramova, Henrik Lehrmann Christiansen and Villy Bæk Iversen

Department of Photonics Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark
Email: {apop, hlch, vbiv}@fotonik.dtu.dk

***Abstract*—Cloud Radio Access Network (C-RAN) is a new mobile radio access network design based on centralized and pooled processing. It offers potential cost savings by utilizing the so-called tidal effect due to user mobility in cellular networks. This paper provides a quantitative analysis of the performance (multiplexing) gain of such cellular networks. The used analytical model is based on a multi-dimensional loss system and can be applied to heterogeneous networks with various cell traffic profiles. Based on the analysis, the key parameters for cell deployment optimization are identified. The conditions for optimization are based on the aggregated traffic characteristics and baseband unit pool dimensioning. This paper considers cells with different traffic profiles and the optimal conditions for maximization of the pooling gain are determined. Furthermore, it is shown how the model can be applied to dynamically re-assign cells to a pool of baseband units. The re-assignment is based on the cell load and traffic characteristics such that effective utilization of the baseband resources is assured.**

Keywords–C-RAN, deployment optimization, multiplexing gain, baseband unit pool dimensioning, multi-dimensional loss system.

I. INTRODUCTION

The explosive increase in mobile traffic is a main driver for a spectrum, energy, and cost efficient design of the future radio access network (RAN). Network densification is a prevailing technique that addresses the challenge of 1000-fold traffic growth of mobile data. The full benefits of network densification can be realized if it is followed by complementary backhaul technology [1], such as Cloud RAN (C-RAN). C-RAN is a scalable and flexible RAN design where the baseband processing is virtualized, centralized and shared among base stations (BS). The centralization of the processing power enables high cooperation among distributed antennas. Virtualization on the other hand allows for processing aggregation and dynamic resource allocation. Thus, C-RAN reduces the operators capital and operating expenditures, provides high spectral and energy efficiency. C-RAN supports coexistence of multi-standard types of communication (device to device, full duplex), and multi-layer architectures. Additionally, C-RAN facilitates the deployment of services at the edge, opens new opportunities for services in the cloud, such as the ability to offer the radio access network as a service [2].

The C-RAN architecture consists of three main parts: remote radio heads (RRHs) that provide the wireless coverage, baseband unit pool (pool of virtual BSs) and a transport network (fronthaul) that connects the BBU pool with the RRHs. The up to date research confirms that the C-RAN design simplifies and reduces the cost of dense cell deployment [3]. Yet the conditions for optimal deployment under C-RAN remain an important area of research. The need for analysis, design and optimization of fronthaul and backhaul technologies for 5G is emphasized in a recent draft proposal of the pre-structuring model for the Horizon 2020 5G Infrastructure PPP [4]. In this work, traffic engineering approach is used in order to perform a quantitative study of C-RAN, and indicate the conditions for optimal multiplexing gain and dimensioning of the BBU pools. The model used in this paper is generalized and can be used for heterogeneous network deployments under various traffic models. The goal of this paper is to determine the key performance metrics that maximize the multiplexing gain. Furthermore, in our model, the optimal dimensioning of the BBU pool considers both the cost saving factor as well as the sensitivity to traffic variations. As a baseline, we consider a network consisting of two cell types that generate different traffic profiles. The work suggests the optimal ratio of the two types of cells for an energy efficient BBU pool, and how the architecture can adapt to the changes in the traffic conditions.

The remainder of the paper is organized as follows: Section II provides an overview of related works. Section III presents the model based on direct routing and how it is mapped to the C-RAN architecture. Section IV discusses the approach taken in this paper for evaluation of the multiplexing gain and dimensioning of the BBU pool. Section V presents and analyses the results for a specific case with respect to multiplexing gain and dimensioning, and elaborates how the model can be applied for dynamic mapping between RRH and BBU pools. Finally, the last section concludes the paper.

II. RELATED WORK

As indicated in [5] the main multiplexing gain in C-RAN comes from the fact that the cells have diverse traffic load during day hours depending on the area they serve. This is the so-called ”tidal effect” since the load in the mobile network moves according to the daily routine of the users. During the working hours more users are located in the office areas, hence the BSs associated to those cells are busiest. After working hours, the users move towards the entertainment and residential areas, increasing the traffic demand on the BSs associated to these cells. In case of traditional deployment, the residential cells during working hours and the office cells during evening hours will be underutilized. The benefit of dynamic assignment of baseband processing to RRHs (illustrated in Figure 1) has been analyzed in [6] through a system level simulation of a scenario where the generated traffic pattern follows the tidal effect. The paper shows that the multiplexing gain comes not only from the fact that the computational power can be shared among

 ISBN: 978-1-61208-411-4

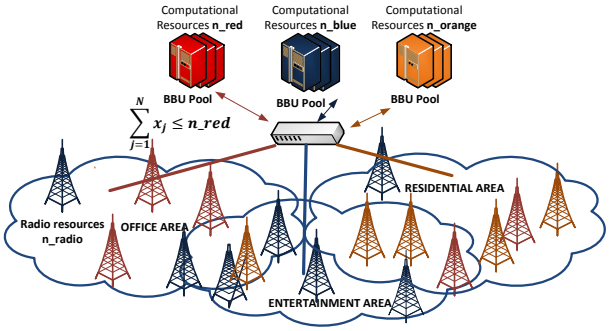


Figure 1. Dynamic allocation of RRHs to a BBU pool. The assignment is defined by different colors.

BSs but also from cost and energy efficiency perspectives. In [7] and [8], the need for dynamic RRH-BBU association is emphasized. Their work shows that the configuration in the network must be flexible in order to provide high performance and energy efficiency. Semi-static and dynamic RRH-BBU switching schemes have been proposed and analyzed with respect to efficiency in the BBU pool. The results show that a percentage of BBUs can be reduced, depending on the traffic load and the applied scheme for assignment.

In [9], the authors model the dynamics of the BBU pool with a multi-dimensional Markov model. The work shows that the system parameters such as pool size, QoS requirements at the radio part, and the traffic load have impact on the system design. In their analysis, all the cells that are associated in a common pool of BBUs, have the same characteristics: size (BS transmission power), type of traffic, QoS demand. Therefore, the proposed model cannot be directly used if heterogeneous deployments are analyzed. In this paper we present a model that can be used to estimate the performance metrics for a C-RAN architecture that can include cells with different size (as number of radio resources), cells with different traffic profiles (smooth, bursty and random), services that have different QoS requirements (as a minimum number of resources that need to be allocated), as well as multi-layer deployment (small cells overlapped with macro cell). The following performance metrics have been studied: blocking probabilities and carried traffic. Based on the desired resource utilization, we dimension the pool of BBUs using the Moe's principle for network dimensioning. We evaluate the dimensioning of the backhaul link based on the carried traffic characteristics. The network model considered in the numerical analysis is based on a mixture of residential and office cells. For the considered scenario, a method is proposed for determining the optimal ratio of the two cell types for multiplexing gain maximization.

III. NETWORK MODEL

This section presents the mathematical model used to asses the benefit of placing baseband processing in a pool that can be shared among RRHs. First, the direct routing network model based on the multi-dimensional systems is described. Afterwards, the mapping of the model to the three different scenarios of network layout is explained.

A. Link model

In a multi-dimensional system, a single link with capacity of n basic units (BUs) is shared among N statistically independent (uncorrelated) flows of Binomial, Poisson, and Pascal (BPP) traffic. A stream is characterized by mean value A_j (offered traffic in number of BUs), standard deviation std_j, the required number of BUs for the entire connection d_j, and n_j is the maximum number of BUs that can be occupied by flow j. The system state at any time can be described by the vector $(x_1, x_2, ..., x_N)$ where $x_j = i_j \cdot d_j$ and i_j represent the number of connections of a flow j. Then the restrictions that lead to truncation of the state space can be formulated as:

$$0 \le x_j \le n_j, \quad \sum_{j=1}^{N} x_j \le n, \quad where \quad \sum_{j=1}^{N} n_j \ge n \qquad (1)$$

In the case where the last two restrictions are not valid (n is sufficiently large such that there is no global restriction), the system corresponds to N independent one-dimensional loss systems (classical BPP loss system), that are represented by state probabilities $p_j(x_j)$.

The system described above is reversible and has product form. Due to the product form, the algorithm based on convolution [10] can be applied to obtain the individual performance metrics of each stream. By successive convolution of one flow at a time, the state probabilities can be aggregated and a one-dimensional vector can be used to describe the system (* denotes the convolution operation):

$$p(x) = p_1(x_1) * p_2(x_2) * ... * p_N(x_N), \qquad (2)$$

where $x = x_1 + x_2 + ... + x_N$. The convolution is done such that first two flows j and k are convolved with limitation $\min(n_j + n_k, n)$. Then the third flow is added to the previous convolution and so on. Due to the truncation, normalization at each step needs to be performed in order to get the true state probabilities. To calculate the time, call, and traffic congestion for a flow j, all flows except j need to be convolved into $p_{N/j}$. The derivation of the three types of congestion is given in [11], here only the calculation for the carried traffic (in number of BUs) is presented:

$$Y_j^n = \sum_{x=0}^{n} \sum_{x_j=0}^{x} x_j \cdot p_{N/j}(x - xj) \cdot p_j(x_j) \qquad (3)$$

and $C_j^n = (A_j - Y_j^n)/A_j$ represents the traffic congestion. By applying the above method, the performance measures for each flow can be derived.

B. Network with direct routing

A network with direct routing [12] is characterized by routes R_j representing different traffic flows, links L_k and $d_{j,k}$ as the number of BUs a route j uses on a link k. Each link is represented with capacity l_k that defines the maximum number of basic unit that all flows can use on that link. The restriction on each link can be expressed as:

$$\sum_{j=1}^{N} x_{j,k} = \sum_{j=1}^{N} i_j \cdot d_{j,k} \le l_k, \quad k = 1, 2, ...K \qquad (4)$$

All the routes are independent, hence the convolution algorithm can be applied to aggregate the state probabilities of any two route to one route, until one route remains for which the performance metrics are calculated. Now, during convolution, each link has to be considered one at a time, as a restriction to the state space. Because each link can restrict one or more

 ISBN: 978-1-61208-411-4

routes, the number of busy channels at each link, or the number of connections at each routes need to be tracked. The algorithm becomes more complex since multi-dimensional vectors need to be convolved, where the number of links defines the dimension. The state number increases to maximum $\prod_{k=1}^{K}(l_k+1)$, which requires large memory for calculation.

C. Network layout mapping to a C-RAN deployment

Using the model presented, the following notation will be used throughout the paper to describe a C-RAN network. A BBU pool is associated with N RRHs, where a RRH j can use up to n_j radio resources. The number of baseband processing power (or computational resources) in a BBU pool is given by n, where $n \leq \sum_1^N n_j$. The traffic at RRH j is represented through the mean value of offered traffic A_j and standard deviation std_j. A call j requires d_j radio and computational resources for the entire duration of a connection. In the multi-dimensional Markov model there will be two types of truncations. The truncation due to the limited radio resources is referred to as blocking probability due to radio resources, while the truncation that is resulted from n is referred to as blocking probability due to computational resources (BBU pool limitation). Hence, for each traffic flow, the call blocking probability depends on the blocking probability due to radio resources and blocking probability due to computational resources.

Using the model with direct routing, the system can be represented through a matrix where the routes are identified as columns, and the links are defined by rows. We consider three different deployment scenarios in C-RAN in order to explain how the analysis can be performed. The reason for this is to show that this method is general and that the complexity of the algorithm can be highly reduced. The reduction can be done both in terms of dimensions of convolution vectors as well as in number of convolutions, due to reduced dependences on the links and generalizations on the cells characteristics.

D. Case study: proportion of office and home small cells

The direct routing equivalent for a network where the BBU pool aggregates a proportion of cells that serve office and residential area is presented in Table I. The number of office RRHs is O, where each RRH has n_o radio resources. The number of residential cells is $N-O$ where each RRH has n_r radio resources. The office cells are offered bursty traffic model with equal mean and standard deviation (Pascal distribution). The traffic at the office cell is modeled using smooth model (Engset distribution) and has equal characteristics among all residential cells. In this paper, this case study is considered as baseline for evaluation of the multiplexing gain in C-RAN. As it can be seen, the table consists of an identity matrix of dimension N. Hence, the complexity of the method described in Section III is highly reduced: the number of the convolutions required to get the performance metrics of one traffic stream is reduced to N. Since there are no dependencies among cells, except the last row, the aggregation of the streams can be done into one-dimensional vectors, and only the global state needs to be remembered. Thus, the number of the states and the required memory is of complexity $O\{n\}$.

E. Case study: mixture of traffic types

This case corresponds to the heterogeneous traffic characteristics in terms of BUs that a stream requires during the connection. Video services that require high bandwidth can be modeled with $d_j > 1$. Table II shows the equivalent direct routing model for C-RAN that aggregates RRH, that offer heterogeneous services in terms of bandwidth demand $d_1 \neq d_2$. The model is for a case of 4 non-overlapping cells, which can be easily extended to more cells. The two traffics types can have individual mean value and standard deviation, while the radio resource limitation could be the same or different. The limitations L_5 and L_6 could be left out, or used when QoS guarantee needs be implemented to make sure that the increase of one type of traffic does not block the other type of traffic. The complexity of the algorithm is again reduced due to the symmetry. The number of convolution for each individual traffic stream (in this case two) is equal to the number of cells, while the dimension of the each convolution vector is equal to the number of different traffic flows. Hence, in the considered example the number of states and the required memory is of complexity $O\{(n_{d1}+1)*(n_{d2}+1)\}$.

TABLE I. Direct routing equivalent to C-RAN that covers a mixture of office and home cells

	Routes							
Link	R_1	R_2	...	R_O	R_{O+1}	..	R_N	Capacity
L_1 ... L_O	Identity matrix of size O				Zero matrix of size $[R, O]$			n_o ... n_o
L_{O+1} ... L_N	Zero matrix of size $[O$ x $R]$				Identity matrix of size R			n_r ... n_r
L_{N+1}	all ones vector of size $[1, N]$							n

TABLE II. Direct routing equivalent to C-RAN that covers a mixture of traffic types for each cell

	Routes								
	Cell1		Cell2		Cell3		Cell4		
Link	R_1	R_2	R_1	R_2	R_1	R_2	R_1	R_2	Capacity
L_1	d_1	d_2	0	0	0	0	0	0	n_r
L_2	0	0	d_1	d_2	0	0	0	0	n_r
L_3	0	0	0	0	d_1	d_2	0	0	n_r
L_4	0	0	0	0	0	0	d_1	d_2	n_r
L_5	d_1	0	d_1	0	d_1	0	d_1	0	n_{d1}
L_6	0	d_2	0	d_2	0	d_2	0	d_2	n_{d2}
L_7	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	n

F. Case study: Multi-layer deployments

Multi-layer heterogeneous deployments are considered as a way of increasing the throughput per area. A scenario where a BBU pool covers cells with different sizes, and traffic offloading exist among overlapping cells, should be considered. The analysis of such a case, should reveal the optimal number of small cells per sector of a macro cell, and could be used to indicate how to dimension BBU pool, depending of the traffic offloaded from the macro cells to the small cells. A direct routing equivalent for a three sector macro cell with two small cells per sector is shown in Table III. All small cells have the same characteristics for the offered traffic and size of a cell (n_m for macro cell and n_s for small cells). The traffic streams in the small cells can also use radio resources in the macro cells, with call rearrangements [12]. Regarding the complexity analysis, this is the most complex case compared to the previous case studies. Two sectors can be easily aggregated into one dimensional vector, so the number of one dimensional convolution is equal to double the number of small cells per sector ($small\_nr\_sector$). In order to find

out the performance metrics for each traffic stream (one for macro cell and one for small cell), the algorithm requires one convolution vector of dimension equal to the number of small cells per sector. Then the number of states increases to order of $(n+1)\cdot(n_m+1)\cdot \prod_{small\_nr\_sector}(n_s+n_m+1)$.

TABLE III. Direct routing equivalent to C-RAN for multi-layer deployment

Links	Routes Sector1			Sector2			Sector3			Capacity
L_1	1	0	0	0	0	0	0	0	0	n_m
L_2	1	1	0	0	0	0	0	0	0	n_m+n_s
L_3	1	0	1	0	0	0	0	0	0	n_m+n_s
L_4	0	0	0	1	0	0	0	0	0	n_m
L_5	0	0	0	1	1	0	0	0	0	n_m+n_s
L_6	0	0	0	1	0	1	0	0	0	n_m+n_s
L_7	0	0	0	0	0	0	1	0	0	n_m
L_8	0	0	0	0	0	0	1	1	0	n_m+n_s
L_9	0	0	0	0	0	0	1	0	1	n_m+n_s
L_{10}	1	1	1	1	1	1	1	1	1	n

IV. Discussion on Multiplexing Gain and BBU Pool Dimensioning

This section outlines the approach considered for evaluation of the multiplexing gain and the conditions for optimal dimensioning and configuration of the pool. The rationals for the considered performance metrics are discussed as well.

A. Multiplexing Gain

In [9], it is demonstrated that, as more cells are aggregated to the BBU pool, the gain (defined as reduction of the number of BBU processing servers that are required to achieve a blocking probability lower than a certain threshold) is increasing. Furthermore, it is shown that as the pool size becomes large, the gain is increasing with a slow pace, such that at a very large pool size, the gain is approaching a limit. Still, the work is missing a discussion on the background for such trend of the gain. The increase in the multiplexing gain comes from the principle of group conservation [13]. In order to explain better, a comparison is made on the n number of resources (BUs) required to achieve a blocking probability of 1% in case of serving individual streams and an aggregation of the N streams. Figure 2 shows the comparison when $N=100$ traffic streams are considered, each with mean value of 10 (offered traffic is 10 erlang) and $std=\sqrt{\sigma^2}=\sqrt{10}$ (Poisson arrivals). The dashed line shows the normalized number of BUs (n/N) when the traffic streams are served independently, which is constant. The full line shows the normalized number of BUs required to serve the aggregated traffic that is decreasing as N is increased until a certain point after which it reaches a limit and becomes almost constant. The reason for this comes from the fact the way the summary statistics are derived for the aggregated stream. Since each stream is independent of the others, the mean and the standard deviation are calculated as:

$$A_{agg}=\sum_{1}^{N}(A_j), \quad std_{agg}=\sqrt{\sum_{j=1}^{N}{std_j}^2} \tag{5}$$

These equations indicate that the mean value of the total traffic is the same in case of individual streams and stream aggregation. The difference is in the standard deviation, or the

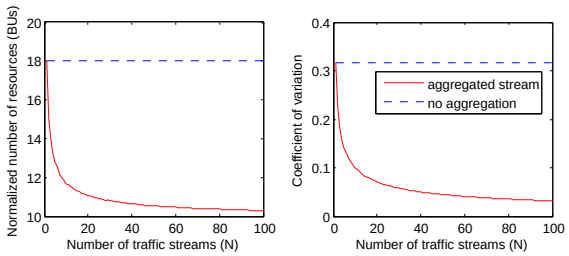


Figure 2. Analysis of multiplexing gain with aggregation.

coefficient of variation ($CV=\frac{std}{A}$) which is shown in Figure 2 to the right. The CV is reduced as the number of streams is increased, but already after $N=30$ the reduction is slow. Any additional increase of N will not lead to significant reduction of the number of required BUs. The channel utilization, defined as A/n will not be significantly improved at large pools, since any marginal increase of the offered traffic will lead to equal increase of number of BUs for each group, meaning $\frac{\Delta A}{\Delta n}$ becomes constant. This means that very large pools will not lead to significant increase of the gain compared to medium size pools. Due to high utilization, very large groups are even more sensitive to overload, and therefore large pools are not recommended. For that reason, the tradeoff between utilization and sensitivity should be considered when dimensioning.

Having in mind the discussion above, the multiplexing gain defined as in (6) is used as a performance metric to evaluate how much the coefficient of variation is reduced in case of the aggregating the individual streams.

$$\text{MultiplexingGain}=\frac{\sum_{j=1}^{N}(A_j+std_j)}{(A_{agg}^{carried}+std_{agg}^{carried})} \tag{6}$$

s.t. $A_{agg}^{carried}$ and $std_{agg}^{carried}$ are carried traffic characteristics.

B. Dimensioning of computational resources.

Two approaches of dimensioning can be considered: dimensioning with fixed blocking probability and dimensioning with fixed improvement function. With fixed blocking probability, the dimensioning of the BBU pools is done by restricting the time congestion to a threshold such that the number of calls that need to re-attempt the connection will be low. This type of dimensioning can easily lead to a system with high utilization (large pool size), but also very sensitive, since it does not consider the channel utilization.

On the other hand, the Moe's principle for dimensioning is based on the improvement function. The improvement function is defined as the increase in carried traffic when the number of channels (n) is increased by 1, $F_n(A)=Y^{n+1}(A)-Y^n(A)$, where $Y^n(A)=\sum_{j=1}^{N}Y_j^n(A_j)$. In this case the point where the $\frac{\Delta A}{\Delta n}$ becomes constant for all BBU pools indicates the dimension of the pools. The improvement function can be set to a fixed improvement value F_{target}, such that balance between high utilization and sensitivity is be ensured. A cost requirement can also be included in determining the optimal number of computational resources. Then the improvement value depends on the cost of the additional resource such as cost of fiber, BUU unit and alternatively the cost of adding a RRH to a BBU pool. The increase of the carried traffic should be included as well as income, such that $F_{target}=\frac{cost}{income}$.

 ISBN: 978-1-61208-411-4

V. C-RAN Optimization

A. Input Parameters

In this work, the study case where a BBU pool aggregates RRH that cover residential and office areas is considered. The chosen parameters for the analysis follow the examples given in [5] and [6]. The total number of cells is $N = 100$, while the percentage of office cells is varied between 1% and 99% with 1% as step. Each cell has $n_r = n_o = 28$ radio resources, which limits the maximum number of computational resources at the BBU pool at $N \cdot n_r = 2800$. The offered traffic, and standard deviation of the office and residential cells are summarized in Table IV. The traffic streams will result in very low radio resource blocking probability. The overall blocking probability will be mostly influenced due to the diagonal truncation which results from the limitation of the resources in the BBU pool. Two sub-cases have been considered as two different time snapshot. One is from daytime when the traffic of the office cell is higher than the traffic from the home cells. The other is in evening time, when the traffic of the residential cells is higher. By considering these two snapshots, the dynamic of the traffic during one day can be captured.

TABLE IV. Input parameters

	Daytime		Evening time	
Cell type	Office	Home	Office	Home
Load	30%	10 %	5%	15%
Traffic type	bursty (Pascal dist.)	smooth (Engset dist.)	smooth (Engset dist.)	bursty (Pascal dist.)
A	8.22	2.77	1.27	4.75
std	3.51	1.44	0.99	2.43

B. Multiplexing gain

The multiplexing gain, according to (6) for the considered case study is shown in Figure 3. During day time the multiplexing gain is reduced as the number of office cell is increased. This is because the mean value is increased but the difference in the standard deviation does not give influence in coefficient of variation of the aggregated and the sum of the individual streams. During night time the opposite trend is observed: the multiplexing gain is increasing as the number of office cells is increased. In this case, the mean value of the aggregation stream is decreasing with the increase of the office cells, and the deviation of the aggregation stream becomes smaller compared to the individual streams.

$$MG = \frac{\sum_{j=1}^{N} \max((A_j + std_j)^{day}, (A_j + std_j)^{night})}{(A_{agg}^{carried} + std_{agg}^{carried})} \quad (7)$$

By looking at the multiplexing gain of the sub-cases of daytime and night time, the optimal ratio of office cells and home cells cannot be deducted. In order to capture the traffic dynamics during one day, (6) has been modified to (7). This metric is also shown in Figure 3 and as it can be seen it peaks at 22% of office cells. Hence, for the this case, the largest gain is achieved when the number of office cell is 22 out of 100.

C. Dimensioning the BBU pool

For dimensioning the BBU pool in terms of computational resources, we use the Moe's principle. We do not focus in this paper on the cost, nor the income. We use improvement

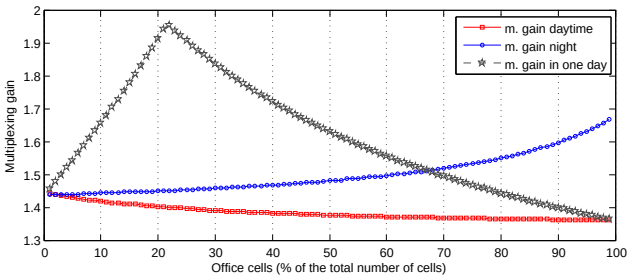


Figure 3. Multiplexing gain according to (6) and (7).

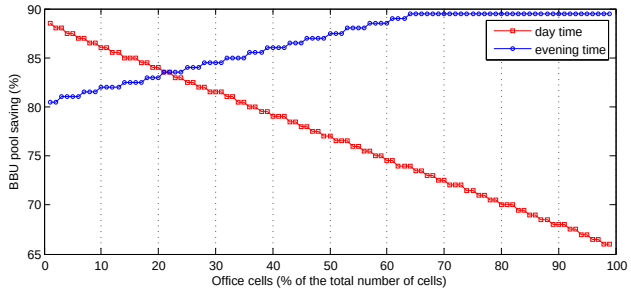


Figure 4. Optimal dimensioning of BBU pool.

value $F_{target} = 0.2$, such that $F_{n-1}(A) > F_{target} \geq F_n(A)$. The analysis has been done for the two considered sub-cases: daytime and night. Instead of giving the optimal number of computational resources, we indicate the percentage of the maximum number of computational resources that can be saved. Figure 4 shows the computational resources percentage that can be saved in case of multiplexing. In daytime analysis, the percentage of the saved computational resources is reduced with the increase of the number of the office cells. The reason for this is that the number of the computation resources scales with the mean value of aggregated traffic. As the mean value of the office cells traffic is larger than the mean value of the home cell traffic, by increasing the number of the office cells, the mean value of the aggregated traffic is increased. During evening time the opposite trend is observed: the increase of the percentage of the office cell reduces the mean value of the aggregated stream, and therefore less computational resources are required. From the figure, it can be noticed that the two lines cross at 22% of the office cells, meaning that with this ratio of office and residential cells, the same savings can be achieved during day time and night time. Hence, the optimal ratio of the office and residential cell is 22 office and 78 residential cells, by which almost 85% of the maximum resources in the pool can be saved. The analysis based on multiplexing gain and dimensioning on the BBU pool has shown the same results. Furthermore, the conclusion is comparable with the simulation based analysis in [6], which confirms the correctness of the described model.

D. RRH-BBU pool dynamic mapping

The optimal percentage of office cells for different mean values of the traffic streams for office and residential cells during day time and during night time is summarized in Figure 5. Additionally for each optimal deployment it shows the potential savings by dimensioning the size of the pool using the Moe's principle. The results show that in case of a change of the traffic characteristics, the model can be used for

 ISBN: 978-1-61208-411-4

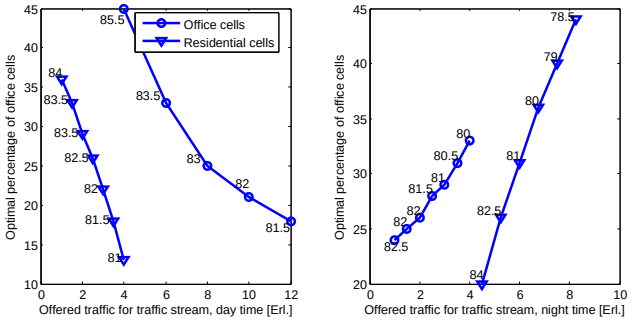


Figure 5. Optimal deployment for variable load during day time.

flexible and dynamic re-assignment of RRH to BBU pools. For example, if the mean value of the traffic stream for residential cells during night time is increased, the number of office cells per BBU pool need to be increased. On the other hand, if the mean value of the residential traffic stream during day time is increased, then the number of office cells need to be reduced.

The radio resourcce blocking probability is low as the load of the cell is not high (Table IV) and the overall blocking probability is influenced from the blocking probability due to computational resources. This is important as the model complexity is further reduced, as only the global state needs to be remembered, which can be described with one dimensional vector of length n. This simple analysis allows for adoption to the dynamic changes in the configuration. If a certain cell needs to be added to the BBU pool, a convolution needs to be performed in order to aggregate the new cell traffic. If one cell needs to be removed, deconvolution needs to be done.

The challenge of the fronthaul design is not just limited to high capacity requirement, but also to the ability to provide flexible and adaptive deployments with respect to RRH-BBU pool assignment. Fiber solutions are capable of supporting high data rates, but are lacking the ability for flexible re-assignment due to the need of manual configurations or very costly optical switches. Adopting any other transport solutions (ex. packet based: wired or wireless) is challenged with strict jitter and synchronization requirements but are capable of flexible reconfigurations. As C-RAN already integrates the concepts of network function virtualization and network virtualizaion ([14], [15]), adoption of software defined networking (SDN) can further optimize and simplify network design and operation. The proposed model can be implemented at an SDN controller. The SDN controller will be responsible for RRH to BBU pool re-assignment due to traffic distribution change and/or addition of new cells in the network. Thus, the SDN controller can instruct and manage all virtual network components in order to maximize the multiplexing gain and dimension the BBU pools optimally. Figure 1 illustrates the dynamic assignment of RRH to BBU pools, where not only the location, but the traffic load and type determine the assignment.

VI. Conclusion

This paper concludes the optimal conditions for dense cell deployments under which the multiplexing gain is maximized. In the presented study case, this is defined as the optimal ratio of the two types of cells: serving office and residential areas. The model has been compared with simulation based analysis, which confirms the correctness of the model. Additionally, we demonstrate that the model indicates the optimal ratio of the cell types depending on the individual traffic loads.

Furthermore, the analysis shows that not only cost, but sensitivity to traffic variations need to be considered when dimensioning the pool of baseband units. For the given ratio of the cell types, the indicated dimension is proven to be optimal.

The model used in the analysis is generalized, and various case studies have been identified. These studies include heterogeneous deployments and different traffic profiles. Due to its simplicity and low level of complexity, we show that the model can be adopted for dynamic re-assignment of RRH to BBU pool. In the future, additional cases are going to be studied, as well as further analysis will be conducted to investigate the implications of new radio technologies such as coordinated multipoint and carrier aggregation.

VII. Acknowledgment

This work was partially sponsored by the 7th Framework Programme for Research of the European Commission HARP project, under grant number HARP-318489.

References

[1] N. Bhushan, et al., “Network Densification: The Dominant Theme For Wireless Evolution into 5G,” Communications Magazine, IEEE, vol. 52, no. 2, February 2014, pp. 82–89.

[2] S. Ferreira, et al., “An architecture to offer cloud-based radio access network as a service,” in Networks and Communications (EuCNC), 2014 European Conference on, June 2014, pp. 1–5.

[3] A. Checko, et al., “Cloud RAN for Mobile Networks - A Technology Overview,” Communications Surveys Tutorials, IEEE, vol. 17, no. 1, Firstquarter 2015, pp. 405–426.

[4] “EC H2020 5G Infrastructure PPP Pre-structuring Model RTD and INNO Strands,” 2014, URL: http://5g-ppp.eu/wp-content/uploads/2014/03/March-2014-_5G-Infra-PPP_Pre-structuringModel_v1-0.pdf [accessed: 2014-12-01].

[5] “C-RAN The Road Towards Green RAN,” China Mobile Research Institute, Tech. Rep., 2011.

[6] A. Checko, A. Checko, H. Holm, and H. Christiansen, “Optimizing Small Cell Deployment by the Use of C-RANs,” in Proceedings of 20th European Wireless Conference, May 2014, pp. 1–6.

[7] C. Liu, K. Sundaresan, M. Jiang, S. Rangarajan, and G.-K. Chang, “The case for re-configurable backhaul in cloud-RAN based small cell networks,” in IEEE INFOCOM, April 2013, pp. 1124–1132.

[8] S. Namba, T. Warabino, and S. Kaneko, “BBU-RRH Switching Schemes for Centralized RAN,” in 7th International ICST Conference on Communications and Networking in China, Aug 2012, pp. 762–766.

[9] J. Liu, S. Zhou, J. Gong, Z. Niu, and S. Xu, “On the Statistical Multiplexing Gain of Virtual Base Station Pools,” in Global Communications Conference (GLOBECOM), 2014 IEEE, Dec 2014, pp. 2283–2288.

[10] V. B. Iversen, “The Exact Evaluation of Multi-Service Loss Systems with Access Control,” Teleteknik, English ed., vol. 31, no. 1, Firstquarter 1987, pp. 56–61.

[11] V. B. Iversen, Teletraffic Engineering. Chapter 7: Multi-dimensional loss systems. Technical University of Denmark, 2013.

[12] V. B. Iversen, V. Benetis, and P. D. Hansen, “Performance of Hierarchical Cellular Networks with Overlapping Cells,” in Proc. EuroNGI Workshop, 2004, pp. 7–19.

[13] M. Stasiak, M. Glabowski, A. Wisniewski, and P. Zwierzykowski, Modelling and Dimensioning of Mobile Wireless Networks: From GSM to LTE, 1st ed. Wiley Publishing, 2011.

[14] R. Wang, H. Hu, and X. Yang, “Potentials and Challenges of C-RAN Supporting Multi-RATs Toward 5G Mobile Networks,” Access, IEEE, vol. 2, 2014, pp. 1187–1195.

[15] C.-L. I, J. Huang, R. Duan, C. Cui, J. Jiang, and L. Li, “Recent Progress on C-RAN Centralization and Cloudification,” IEEE Access, vol. 2, 2014, pp. 1030–1039.

 ISBN: 978-1-61208-411-4

False Alarm Rate Analysis of the FCME Algorithm in Cognitive Radio Applications

Johanna Vartiainen and Risto Vuohtoniemi

Centre for Wireless Communications
University of Oulu
Oulu, Finland
Email: firstname.lastname@ee.oulu.fi

***Abstract*—Cognitive radio is a promising choice to fulfill needs of growing wireless applications in the future. Spectrum sensing is beneficial in several circumstances when secondary user (SU) search empty frequencies for a transmission. One interesting choice for spectrum sensing is the localization algorithm based on the double-thresholding (LAD) method. The LAD method is based on the forward consecutive mean excision (FCME) algorithm that calculates the used thresholds. Threshold setting is based on the usage of the desired false alarm rate, which is sensitive to issues like the length of the integration time. In the real-time applications, integration time is limited. In this paper, the false alarm rate of the FCME algorithm is studied. The false alarm rates of the FCME algorithm in the noise-only case with different integration times (sample vector lengths) are analyzed. The minimum length of the sample vector is defined. The simulation results are verified by the real measurement results in the noise-only case, and a scenario that combines the results is presented. It is also noted that in the noise measurements, the achieved false alarm rates are somewhat lower than the desired ones.**

Keywords–cognitive radio; spectrum sensing; false alarm rate.

I. INTRODUCTION

Cognitive radio technology [1] [2] [3] can be considered as a revolution against the traditional, inflexible frequency allocation. Cognitive radio (CR) enables both dynamic spectrum management and flexible transmission bandwidth [4]. In CR, secondary users (SU) may transmit if there is room aka empty frequencies (white space) in the spectrum and if they are not interfering primary users (PU). Interference-free transmission is a privilege of the PUs. In cognitive radio, SUs may find out empty frequencies using, for example, databases or spectrum sensing [5] [6] [7] [8]. Sensing may be beneficial instead of geolocation and databases, for example, in the wireless local area network (WLAN)-type solutions when transmitters are located close to each other and transmit powers are small. Also, public safety applications when the connection to the outside world is lost may use sensing. Spectrum sensing requires ability to find unused frequencies, which can be done via detecting existent signals.

Many detection methods are based on the use of a threshold. The basic principle is that the threshold separates the samples into two sets: noise and signal sets. Nowadays, most of the methods use adaptive thresholds. Threshold setting is a very demanding task, especially when the threshold is set adaptively. As too high threshold causes missed detections, too low a threshold leads to false detections. Missed detection means that existing signals are not detected, as false detection means that noise samples are falsely detected to be from a signal.

One of these detection methods is the localization algorithm based on the double-thresholding (LAD) [9]. At the core of the LAD method, the forward consecutive mean excision (FCME) algorithm [10] provides the used detection thresholds. The FCME algorithm sets the threshold iteratively based on the mean of sample energies and a pre-selected threshold parameter. This parameter depends on the statistical properties of the noise-only case. Usually, Gaussian assumption is used even though the measured noise is not purely Gaussian [9]. The threshold parameter is defined using the desired false alarm rate $P_{FA,DES}$. It defines how many samples are above the threshold when there is only noise present. The FCME method uses constant false alarm rate (CFAR) principle, so the false alarm probability stays almost constant. However, it is sensitive to the issues like the length of the considered sample vector and noise properties. In the real-time applications, integration time is limited, so the number of considered samples N can not be as large as in the computer simulations.

The performance of the FCME algorithm is highly depending on the false alarm rate. If the achieved false alarm rate differs from the desired one, the performance of the FCME method may degrade. Especially when the signal-to-noise ratio (SNR) is low, the false alarm rate totally defines the performance of the FCME algorithm, and, thus, the LAD method. If the achieved false alarm rate is not close enough to the desired one, the performance of the LAD method may even totally degrade. It is very important to control $P_{FA,DES}$ because it is related to the caused interference as well as the spectrum opportunity loss in cognitive radio applications [9]. Thus, it is very important to study and analyze the false alarm rate of the FCME algorithm.

In this paper, the false alarm rate of the FCME algorithm is studied in the noise-only case. That is, there are no signals present. First, the effect of the length of the considered sample vector (i.e., integration time) to the false alarm rate of the FCME algorithm is analyzed using simulation software generated AWGN noise. Mean, variance as well as minimum and maximum values of achieved false alarm rates are analyzed. Based on those, proper sample vector lengths are recommended. The analysis results are verified by the real measurement results in the noise-only case. The measurements covering a wide range of the spectrum are used to find out the differences in the achieved false alarm rate between the measured and simulation software generated noise. Several measurements up to 39 GHz are used to cover higher frequency areas possible used in future applications as 5G and beyond. The Kruskal-Wallis test is used to provide more statistical information. In addition, a scenario that combines the analysis and measurement results is presented.

 ISBN: 978-1-61208-411-4

This paper is organized as follows. In Section II, the used FCME algorithm is presented. Section III covers the probability of false alarm analysis of the FCME algorithm, and Section IV describes our scenario. Conclusions are drawn at Section V.

II. THE FCME ALGORITHM

The FCME algorithm [9] [10] [11] was originally proposed for impulsive interference suppression in the time domain. Later on, it was noticed that the method can be used also in other transform domains, e.g., in the frequency domain. Its enhanced version called the LAD method [9], which uses the FCME thresholds was developed to detect narrowband information signals, e.g., for spectrum sensing purposes.

The FCME algorithm is blind and independent of modulation methods, signal types and number of signals. The only requirements are that the signal(s) can not cover the whole bandwidth under consideration, and the signal(s) are above the noise level.

The FCME algorithm is computationally simple but effective. It calculates the threshold iteratively based on the noise properties.

Initial Preparation: When the noise is assumed to be zero mean, independent, i.i.d. Gaussian noise, i.e., samples x_i follow the Gaussian distribution, the FCME algorithm calculates the threshold parameter based on [10]

$$T_{CME} = -ln(P_{FA,DES}), \quad (1)$$

where $P_{FA,DES}$ is the desired clean sample rejection rate (the desired false alarm rate) [10]. For example, if the desired clean sample rejection rate is 1% $(= 0.01)$, $T_{CME} = 4.6$ [9]. After that, energy of samples is calculated. Now, samples $|x_i|^2$ that follow the chi-squared distribution with two degrees of freedom are rearranged in an ascending order according to their sample energy. Then, $m = 10\%$ of smallest samples are selected to form the initial set Q (called also as a "clean set").

Algorithm: The FCME threshold is [10]

$$T_h = T_{CME}\overline{Q}, \quad (2)$$

where $\overline{Q}$ denotes the mean of Q. Samples below the threshold are added to the set Q and new mean and threshold are calculated. This is repeated until there are no new samples below the threshold. Usually, it takes 3-4 iterations to get the final threshold. In the end, samples *above* the threshold are assumed to be signal samples, as samples *below* the threshold are assumed to be noise samples.

The required false alarm rate $P_{FA,DES}$ is related to the threshold. Small $P_{FA,DES}$ value leads to larger threshold. Thus, the amount of false alarms is small. Large $P_{FA,DES}$ value leads to smaller threshold and the amount of false alarms is larger [12]. In cognitive radio applications, it is important to control $P_{FA,DES}$ because it is related to the caused interference as well as the spectrum opportunity loss [9].

It should be noted that (1) is valid when the noise is at least approximately Gaussian. It is also possible to define the used equation to other distributions [9]. Note, that the noise variance has no influence [13].

TABLE I. ACHIEVED P_{FA} WHEN $P_{FA,DES} = 0.01$.

N	mean(P_{FA})	diff	var(P_{FA})	min	max
64	0.025245	0.0152	0.0075937	0	0.9062
128	0.015415	0.0054	0.00062043	0	0.8984
256	0.0141	0.0041	$7.505e-05$	0	0.0585
512	0.013526	0.0035	$3.566e-05$	0	0.0390
1024	0.013313	0.0033	$1.721e-05$	0.00195	0.0341
2048	0.013181	0.0032	$8.356e-06$	0.00341	0.0268
4096	0.013139	0.0031	$4.318e-06$	0.00659	0.0229

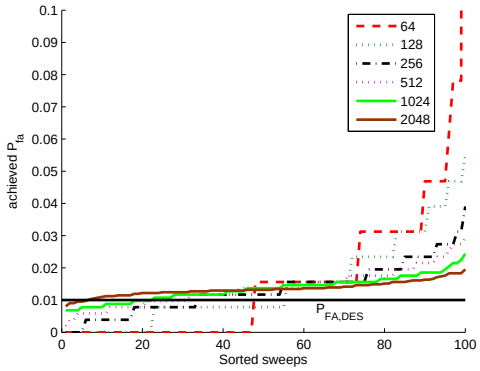


Figure 1. Achieved P_{FA} values for different values of N. MC=100 sweeps. Results are sorted in an ascending order. Matlab-generated noise. $P_{FA,DES} = 0.01$.

III. P_{FA} ANALYSIS OF THE FCME ALGORITHM

Achieved P_{FA} values for different desired $P_{FA,DES}$ values were studied. That is, how close the achieved P_{FA} values are to the desired $P_{FA,DES}$ value. This effects to the performance of the FCME method, especially at low SNR values. Two different commonly used desired $P_{FA,DES}$ values were used, $0.01 = 1\%$ ($T_{CME} = 4.6$) and $0.001 = 0.1\%$ ($T_{CME} = 6.9$) [9]. It means that according to the CFAR principle, when there is only noise present, 1% or 0.1% of the samples should be above the threshold, respectively. In the computer simulations, the effect of the length of the samples N, was considered. The purpose was to find the smallest N when the FCME algorithm is able to operate properly. Measurement results are compared to the Matlab-generated AWGN noise results.

A. Matlab simulations

In the simulations, Matlab software was used. Computer-generated AWGN noise was used as a noise. There were 10 000 Monte Carlo iterations. The length of the samples, N, varied. This is because in the simulations we can use large values of N, but in the real-time implementations, N may be often smaller because of hardware limitations. The purpose was to find smallest N so that the achieved P_{FA} values are in the decent level.

In Table I, achieved P_{FA} values when desired $P_{FA,DES} = 0.01 = 1\%$ and N varies are presented. Diff=$|P_{FA,DES} - P_{FA}|$. As can be seen, means are close to each others when N is large enough, that is, $N >= 256$. Achieved P_{FA} values differ from the desired $P_{FA,DES}$ value 152% ($N = 64$), 54% ($N = 128$), 41% ($N = 256$), 33% ($N = 512$), 33% ($N = 1024$), 31% ($N = 2014$), and 31% ($N = 4096$). It can also

 ISBN: 978-1-61208-411-4

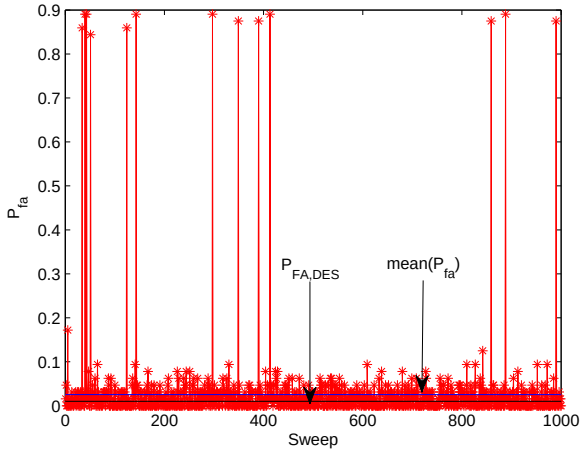


Figure 2. Achieved P_{FA} values when $N = 64$. MC=1000, $P_{FA,DES} = 0.01$.

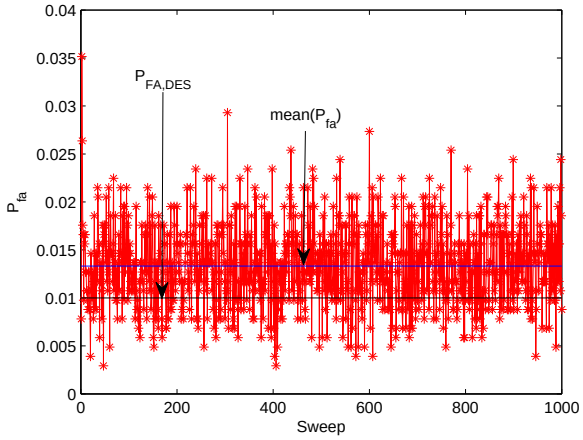


Figure 3. Achieved P_{FA} values when $N = 1024$. MC=1000, $P_{FA,DES} = 0.01$.

been seen that the smaller N (the shorter data), the higher the variance is.

In Figure 1, achieved P_{FA} values are presented for different values of N. There were 100 iterations (sweeps) and the results were sorted in an ascending order. Horizontal line presents desired $P_{FA,DES}$ value. It can be seen that the more samples, the closer the achieved P_{FA} values stay with the desired $P_{FA,DES}$ value (here, $P_{FA,DES} = 0.01$).

In Figures 2 - 4, achieved P_{FA} values are presented when $N = 64$, 1024 and 4096. It can be noticed that the achieved mean of P_{FA} is slightly higher than the desired $P_{FA,DES}$ value. It can also be seen that when N is small (Figure 2), variance is very high.

In Figure 5, variances of the achieved P_{FA} values are considered as in Figure 6, mean P_{FA}, min P_{FA} and max P_{FA} values are studied. In both figures, N varies. It can be seen that when $N >= 256$, values are on acceptable level.

In Table II, achieved P_{FA} values when desired $P_{FA,DES} = 0.001 = 0.1\%$ and N varies is presented. Diff=$|P_{FA,DES} - P_{FA}|$. $P_{FA,DES} = 0.1\%$ means that when $N = 1000$, 1 sample is above the threshold. Thus, we considered $N >= 1024$ to get realistic results; therefore, smaller values for N were not considered.

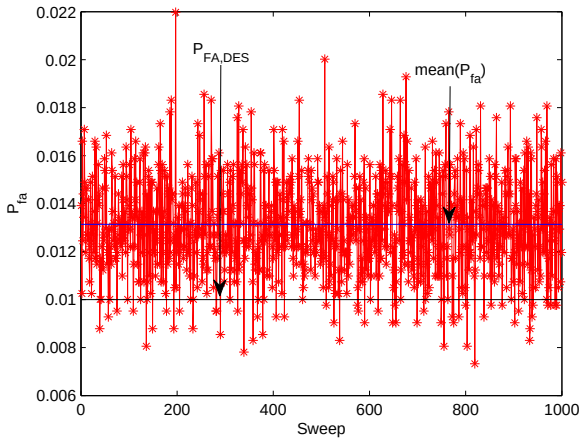


Figure 4. Achieved P_{FA} values when $N = 4096$. MC=1000, $P_{FA,DES} = 0.01$.

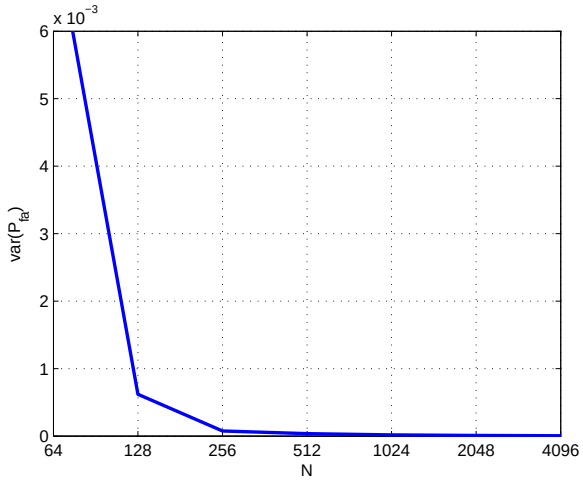


Figure 5. Variance of P_{FA} values for different sample lengths N. Matlab-generated noise. $P_{FA,DES} = 0.01$.

TABLE II. ACHIEVED P_{FA} WHEN $P_{FA,DES} = 0.001$.

N	mean(P_{FA})	diff	var(P_{FA})	min	max
1024	0.0010574	$5.74e-05$	$1.1071e-06$	0	0.00683
2048	0.0010685	$6.85e-05$	$5.5935e-07$	0	0.00585
4096	0.0010708	$7.08e-05$	$2.7042e-07$	0	0.00341

B. Measurements at 10 MHz-39.1 GHz

The measurements were performed in wide frequency area to get reliable and wide-ranging results. Here, high-performance spectrum analyzer (Agilent E4446A) [14] was used as in [15]. Note, that the results depend on the used equipment. We used Instrument Control Toolbox to connect Matlab to the spectrum analyzer to enable direct results

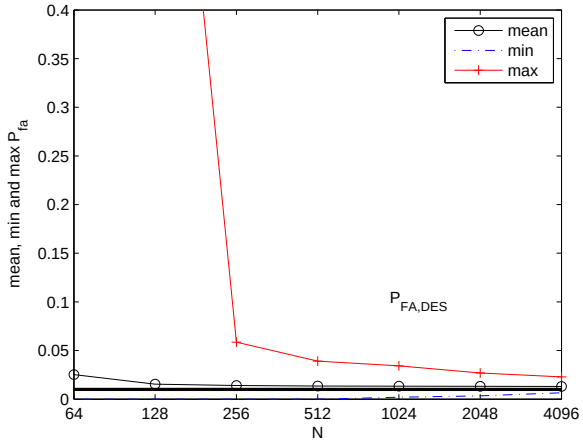


Figure 6. Mean P_{FA}, min P_{FA} and max P_{FA} values for different sample lengths N. Matlab-generated noise. $P_{FA,DES} = 0.01$.

TABLE III. ACHIEVED P_{FA} WHEN $P_{FA,DES} = 0.01$.

freq. range	mean(P_{FA})	var(P_{FA})	min	max
$10 - 110$ MHz	0.006431	$5.8612e-06$	0	0.0149
$1 - 1.1$ GHz	0.0061711	$5.6578e-06$	0.000624	0.0181
$2.5 - 2.6$ GHz	0.0070012	$1.1744e-05$	0	0.0231
$9 - 9.1$ GHz	0.0070244	$1.2441e-05$	0	0.0199
$17 - 17.1$ GHz	0.0060668	$5.5949e-06$	0.000624	0.0149
$39 - 39.1$ GHz	0.0071974	$1.103e-05$	0	0.0199
Matlab-noise	0.013229	$9.7771e-06$	0.000624	0.0237

analysis. At frequency ranges 10-110 MHz, 1-1.1 GHz, 17-17.1 GHz and 39-39.1 GHz, only internal noise level was measured. In frequency ranges 2.5-2.6 GHz and 9-9.1 GHz, broadband antenna was connected, so the noise consists of internal noise and noise from antenna. There were 1 000 time domain sweeps and $N = 1601$ frequency points [15]. Energy of the samples was measured in the frequency domain. Matlab-generated AWGN noise with same N was used for a comparison.

In Table III, achieved P_{FA} values when desired $P_{FA,DES} = 0.01 = 1\%$ and $N = 1601$ are presented. It can be noticed that mean P_{FA} values are very close to each others. Variances are on the same level. It should be noted that now the achieved P_{FA} values are slightly lower than desired $P_{FA,DES}$ value.

In Figure 7, achieved P_{FA} values are presented for different measured frequency bands. There were 1000 iterations (sweeps) and the results were sorted in an ascending order. Horizontal line presents desired $P_{FA,DES}$ value (=0.01). Matlab-generated noise results are presented as a reference. Here, $N = 1601$. It can be seen that the measured results are almost on the same level, and lower than the reference results.

In Figure 8, variances of the achieved P_{FA} values are considered as in Figure 9, mean P_{FA}, min P_{FA} and max P_{FA} values are studied for different measured frequency bands. In both figures, $N = 1601$. It can be seen that there are only small differences.

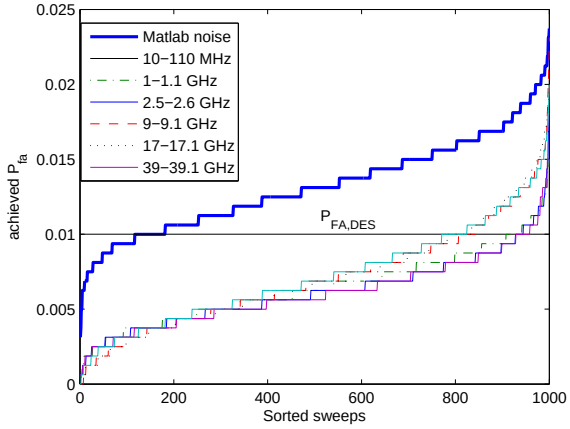


Figure 7. Achieved P_{FA} values for different frequency areas. MC=100 sweeps. Results are sorted in an ascending order. $N = 1601$. Measured noise.

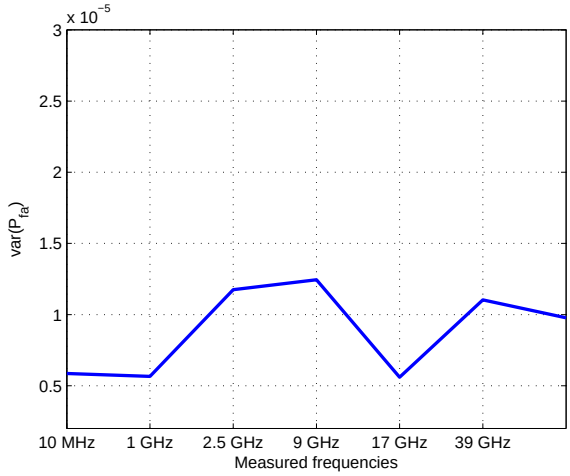


Figure 8. Variance of P_{FA} values for different frequency areas. N=1601. Measured noise.

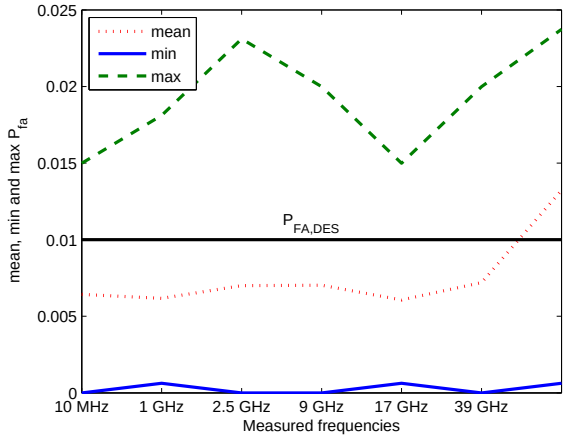


Figure 9. Mean P_{FA}, min P_{FA} and max P_{FA} values for different frequency areas. N=1601. Measured noise.

 ISBN: 978-1-61208-411-4

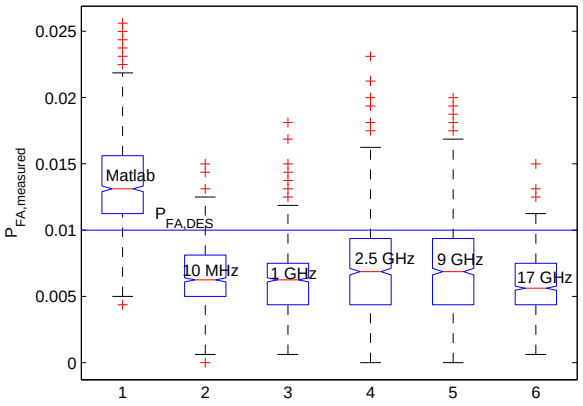


Figure 10. Kruskal-Wallis test to P_{FA} for several measured groups at 10 MHz-17 GHz. $P_{FA,DES} = 0.01$

C. Kruskal-Wallis test

Kruskal-Wallis tests the null hypothesis that samples that are independent and come from two or more groups follow same distribution and their means are equal [16]. There is no normality assumption nor assumptions about the mean and variance. Here, Kruskal-Wallis test is used to produce statistical boxplots.

In Figure 10, Kruskal-Wallis boxplots are presented to achieved P_{FA} for several measured groups at 10 MHz-17 GHz. One boxplot presents five statistics - from bottom to top those are minimum, first quartile, median value (line in the middle of the box), third quartile, and maximum value. This figure confirms the results presented earlier.

IV. SCENARIO

Sensing can be verified using a spectrum analyzer. Here, Agilent E4446A was used, but there are a lot of other equipments, like the wireless open-access research platform (WARP) [17]. The WARP is a platform used to test and prototype wireless networks. The noise level (from internal noise) may vary between the equipments. Therefore, adjusting is needed if it is required that the achieved false alarm rate is controlled. Assume that the LAD method which uses the FCME thresholds is used to perform spectrum sensing. It is desired that the P_{FA} is controlled so spectrum opportunities are not lost. It is possible first to measure the noise in the desired frequency area. As noticed here, the length of the noise vector has to be at least 256 samples when $P_{FA,DES} = 0.01$. It does not matter what is the used sampling rate, however, the same rate should be used later. After measuring the noise level, the FCME threshold can be fixed to correspond the theoretical one. This can be done using a correction coefficient which can be defined when $P_{FA,DES}$ and P_{FA} are known. Note, that this method is valid when the noise is not impulsive.

V. CONCLUSION

The false alarm rate of the FCME algorithm was studied in the noise-only case. A proper length of the sample vector was defined, and analysis results were compared with the results from noise measurements. This result can be used in future simulations and in real-time applications, for example, when implementing the FCME algorithm on the wireless open-access research platform. It was also noted that as in the computer simulations the achieved false alarm rates were slightly higher than the desired ones, in the noise measurements, the achieved false alarm rates were slightly lower than the desired ones. Based on this information, used thresholds can be fixed using a proper correction coefficient in the cases when the achieved false alarm rate need to be as close as the desired false alarm rate as possible. In the computer simulations, the false alarm rate can be reduced as in the measurements and real-time applications, the false alarm rate can be raised.

ACKNOWLEDGMENT

The research of Johanna Vartiainen was funded by the Academy of Finland.

REFERENCES

[1] V. Chakravarthy, A. Shaw, M. Temple, and J. Stephens, “Cognitive radio - an adaptive waveform with spectral sharing capability,” in IEEE Wireless Commun. and Networking Conf., New Orleans, LA, USA, Mar.13–17 2005, pp. 724–729.

[2] S. Haykin, “Cognitive radio: Brain-empowered wireless communications,” IEEE Journal in Selected Areas in Comm., vol. 23, no. 2, Feb. 2005, pp. 201–220.

[3] J. Mitola III and G. Q. Maguire Jr., “Cognitive radio: making software radios more personal,” IEEE Pers. Commun., vol. 6, no. 4, 1999, pp. 13–18.

[4] Z. Feng and Y. Xu, “Cognitive TD -LTE system operating in TV white space in China,” 2013, ITU-R WP 5A, Geneva, Switzerland.

[5] S. Haykin, D. J. Thomson, and J. H. Reed, “Spectrum sensing for cognitive radio,” Proceedings of the IEEE, vol. 97, no. 5, May 2009, pp. 849–877.

[6] C. Liu, M. Li, and M. L. Jin, “Blind energy-based detection for spatial spectrum sensing,” IEEE Wireless Communication Letters, vol. 4, no. 1, Feb. 2015, pp. 98–101.

[7] L. Arienzo and D. Tarchi, “Statistical modeling of spectrum sensing energy in multi-hop cognitive radio networks,” IEEE Signal Processing Letters, vol. 22, no. 3, Mar. 2015, pp. 356–360.

[8] A. Alabbasi, Z. Rezki, and B. Shihada, “Energy efficient resource allocation for cognitive radios: A generalized sensing analysis,” IEEE Transactions on Wireless Communications, vol. PP, no. 99, 2015, pp. 1–15.

[9] J. Vartiainen, “Concentrated signal extraction using consecutive mean excision algorithms,” Ph.D. dissertation, Acta Univ Oul Technica C 368. Faculty of Technology, University of Oulu, Finland, Nov. 2010, {URL}:http://jultika.oulu.fi/Record/isbn978-951-42-6349-1 [accessed: 2015-05-05].

[10] H. Saarnisaari, P. Henttu, and M. Juntti, “Iterative multidimensional impulse detectors for communications based on the classical diagnostic methods,” IEEE Trans. Commun., vol. 53, no. 3, March 2005, pp. 395–398.

[11] H. Saarnisaari and P. Henttu, “Impulse detection and rejection methods for radio systems,” Boston, MA, USA, Oct. 2003, pp. 1126 – 1131, CD-rom.

[12] S. M. Kay, Fundamentals of statistical signal processing: Detection theory. Upple Saddle River, NJ, USA: Prentice Hall, 1998.

[13] H. V. Poor, An introduction to signal detection and estimation, 2nd ed. Berlin, Germany: Springer-Verlag, 1998.

[14] “Agilent,” 2015, URL: http://www.agilent.com [accessed: 2015-05-05].

[15] J. Vartiainen and R. Vuohtoniemi, “Performance of the LAD spectrum sensing method in measured noise at frequency ranges between 10 MHz and 39 GHz,” in AICT, Paris, France, Jul. 2014, pp. 144–149.

[16] W. H. Kruskal and W. A. Wallis, “Use of ranks in one-criterion variance analysis,” J. Amer. Statist. Ass., vol. 47, no. 04, 1952, pp. 583–621.

[17] “WARP,” 2015, URL: http://cmc.rice.edu/warp [accessed: 2015-05-05].

 ISBN: 978-1-61208-411-4

An Efficient Buffer Delay Correction Algorithm to VoIP

Fabio Sakuray
Computer Science Department
University of Londrina
Londrina, Brazil
Email: sakuray@uel.br

Robinson S. V. Hoto
Mathematics Department
University of Londrina
Londrina, Brazil
Email: hoto@uel.br

Gean D. Breda
and Leonardo S. Mendes
Faculty of Electrical and Computing Engineering
University of Campinas
Campinas, Brazil
Email: gdbreda@gmail.com lmendes61@gmail.com

***Abstract*—Audio applications are widely used on the Internet today. In these applications, packets are considered lost if received after their playout time. Such applications require a playout buffer in the receiver for smoothing network delay variations to enable the reconstruction of a periodic form of the transmitted packets. The objective of buffer delay adjustment algorithms (BDA) is to control the packet loss rate using minimum buffer size to jitter smooth. However, current algorithms fail to obtain a particular packet loss percentage. This paper presents a definition of Optimum Buffer Delay (OBD), used to remove jitter and a technique to correct the buffer delay from any BDA applied between talkspurts, with the purpose of bring the packet loss percentage closer to the value defined by audio applications. This new technique is called Buffer Delay Correction Algorithm (BDCA).**

Keywords–Playout Delay; VoIP; Buffer Delay; MOS.

I. INTRODUCTION

Nowadays, the Internet has been broadly used for voice applications, this can be explained by increase in Voice over Internet Protocol (VoIP) applications efficiency and best network bandwidth to users. Unlike of other applications, VoIP can tolerate some packet loss, but none jitter is allowed [1]. In receiver side of VoIP systems, the audio samples must be played as a continuous stream. This is a challenging process because IP present delay variation (or jitter), this phenomenon results in increase on packet loss rate whenever a packet is received after your playout time [2]. The receiver audio application uses a de-jitter buffer that insert an artificial delay (called Buffer Delay) to reduce this effects, resulting in a controlled packet loss rate that enable a greater communication quality. But long buffer delays can reduce the voice quality in interactive audio applications.

To adapt to network delay variations, the buffer delay needs to be continuously changed in order to reduce packet loss rate. The buffer delay control has been studied in many previous works and several Buffer Delay Algorithms (BDA) have been proposed.

However, these BDAs do not produce the packet loss rate as user requested. This paper presents the formal definition of Optimum Buffer Delay (OBD) to jitter remove, and explain how to use this result in Buffer Delay Correction Algorithm (BDCA), a technique to adjust the buffer delay produced by others BDAs. The BDCA has its focus on shaping the packet loss percentage to follow the one defined by voice service while bringing down one-way delay. This work considers only packet loss caused by jitter.

The remainder of this paper is as follows: Section II presents a review of BDAs, Section III details the features of de-jitter buffer, Section IV presents a definition of optimum buffer delay and the BDC) and Section V demonstrates performance comparisons between BDAs. Concluding remarks and future directions are presented in Section VI.

II. BACKGROUND

Figure 1 shows packets sent between two remote VoIP applications in a regular call, where talkspurts are periods with packets transmission and silence are periods without transmission. In a talkspurt k with n^k packets, a packet i is sent at instant t_i^k, received at instant a_i^k and executed in p_i^k (playout time) [3].

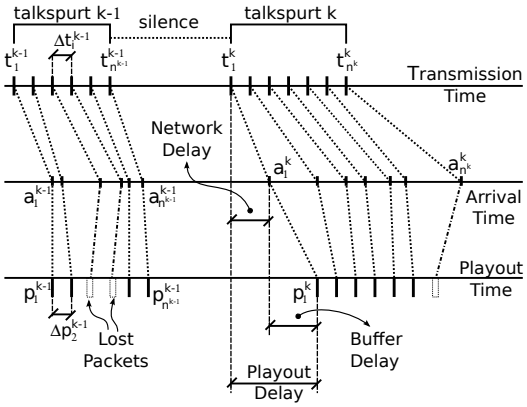


Figure 1. Timings of Packet Audio Transmission.

In the receiver side of VoIP applications, the audio packets must be scheduled to playout with the same temporal spacing used in transmission ($\Delta t_i^k = \Delta p_i^k$). However, jitter makes packets arrive after its playout time and are considered lost because they can not be used when $p_i^k < a_i^k$. To avoid this, most applications use a buffer delay that can be inserted at beginning of each talkspurt (see talkspurt k in Figure 1), which is referred to as "inter-talkspurt" technique, or inserted inside a talkspurt, which is referred to as "intra-talkspurt". This work analyses only algorithms that act in silence periods, since they represent the most of BDA solutions in literature [4].

Lobina in [5], present an important classification of BDAs, as:

1) Packet Loss Intolerant: Algorithms that use high buffer delay values, avoiding packet loss. The sim-

plicity of implementation is the main advantage of these algorithms;
2) Packet Loss Tolerant: audio applications can lose a certain number of packets without affecting audio quality. This class of algorithms adjusts buffer delay to control the packet loss rate;
3) Quality Based: this algorithm class monitors the call quality parameters to adjust the buffer delay.

Another element of voice call is the phenomenon called spike [6], defined as a sudden and large increase in the end-to-end delay. As result the receiver have an interval without packets followed by a series of packets arriving almost simultaneously. Delay spikes represent a serious problem for VoIP applications, since they lead BDAs to overrated buffer delay values. A BDA must react adequately to the spike by changing your behavior.

Several BDAs has been developed with most of them trying to foresee network delay to set the buffer delay. Now let us consider some examples. The next two algorithms are packet loss intolerant. Ramjee in [7] presents four algorithms to measure the delay variance and estimate the average end-to-end delay, the fourth can detect spike and change the algorithm behavior. Barreto and Arago in [8] present an algorithm based on the standard (Box-Jenkins) linear auto-regressive (AR) model. The playout delay estimated ($\hat{d}^k$) of talkspurts k can be write by:

$$\hat{d}^k = \theta_1^\mu \mu(A^{k-1}) + \theta_1^\sigma \sigma(A^{k-1}) + \theta_2^\mu \mu(A^{k-2}) + ... + \theta_n^\sigma \sigma(A^{k-n}) \quad (1)$$

where A^k is network delay of k-th talkspurt; θ_i^μ and θ_i^σ are weights associated with mean ($\mu(A^k)$) and standard deviation ($\sigma(A^k)$), n is the sliding window size with last talkspurts received.

In a call with M talkspurts, (1) can rewrite by $\mathbf{d} = \mathbf{X}\theta$, where matrix $\mathbf{X} \in \mathbf{R}^{M \times 2n}$ is defined as:

$$\mathbf{X} = \begin{bmatrix} \mu(A^n) & \sigma(A^n) & .. & \mu(A^1) & \sigma(A^1) \\ \mu(A^{n-1}) & \sigma(A^{n-1}) & .. & \mu(A^2) & \sigma(A^2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu(A^{M-2}) & \sigma(A^{M-2}) & .. & \mu(A^{M-n-1}) & \sigma(A^{M-n-1}) \\ \mu(A^{M-1}) & \sigma(A^{M-1}) & .. & \mu(A^{M-n}) & \sigma(A^{M-n}) \end{bmatrix}$$

The vectors $\theta \in \mathbb{R}^{2n}$ and $\mathbf{d} \in \mathbb{R}^M$ are: $\theta = [\theta_1^\mu \ \theta_1^\sigma ... \theta_n^\mu \ \theta_n^\sigma]^T$, $\mathbf{d} = [d_{n+1} \ d_{n+2} ... d_{M-1} \ d_M]^T$ with the superscript T denoting matrix transposition.

The estimate of θ is given by $\hat{\theta} = \left[\mathbf{X}^T\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{d}$. However, the matrix $\left[\mathbf{X}^T\mathbf{X}\right]$ may be non-invertible, in which case Barreto and Aragao replace it by its regularized version:

$$\hat{\theta} = \left[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right]^{-1}\mathbf{X}^T\mathbf{d} \quad (2)$$

where $\mathbf{I} \in \mathbb{R}^{2n \times 2n}$ is the identity matrix and $0 < \lambda \ll 1$. The values used by the authors are $\lambda = 0.01$ or $\lambda = 0.001$.

The next three algorithms are packet loss tolerant. Moon *et al.* [3] use the network delay distribution in the last w received packets and a desired packet loss rate. This algorithm can detect spike. Fujimoto *et al.* [9] uses the same idea, but focused on the tail of the network delay probability distribution function. Assuming Pareto distribution for the tail, this approach presents better results when compared with algorithms that use a complete network delay distribution. In [10], Ramos *et al.* present the Move Average Algorithm (MA) to adjusts the playout delay at each new talkspurt given a desired target of average loss percentage (ρ). The authors compute the optimal playout delay (D_k) at the beginning of talkspurt k as:

$$D_k = SORT\left\{Z_i^k\right\} \quad \text{with } i = round(1-\rho)N_k$$

with N^k the number of audio packets received during k-th talkspurt and Z_i^k the variable portion of the end-to-end delay of i-th packet.

The predicted value of D_{k+1}, denoted by $\hat{D}_{k+1}$, is given by

$$\hat{D}_{k+1} = \sum_{l=1}^{M} a_l D_{k-l+1}$$

The coefficients a_l must minimize the mean square error between D_{k+1} and $\hat{D}_{k+1}$. They can by found from solving the equation:

$$\sum_{m=0}^{M} a_{m+1} r_D(m-l) = r_D(l+1) \text{ with } l = 0, 1, ..., M-1.$$

Suppose that it is known the last K values of r_D,

$$r_D(r) \simeq \frac{1}{K-|r|} \sum_{k=1}^{K-|r|} D_k D_{k+|r|}$$

with $r = 0, \pm1, \pm2, ..., \pm(K-1)$. The model's order M is computed as follow: starting with $M = 1$, compute all values of $\hat{D}_k$ and estimate $\mathbb{E}\left[(D_k - \hat{D}_k)^2\right]$, increase M and repeat the process. The model's order is taken equal to the lowest value of M preceding an increase in mean square error.

The next algorithms are quality based. Fujimoto *et al.* [11] shows that jitter, packet loss rate, codec and other parameters can affect call quality. Most solutions only allow packet loss rate setup. The algorithm presented in [11], called E-MOS, utilizes Mean Opinion Score (MOS [1], [12]) classification as input to buffer delay adjust. MOS values are 1 to 5, where 1 is the worst and 5 the best.

Valle *et al.* in [13] present the Dynamic Management of Dejitter Buffer Algorithm (DMDB), that uses MOS rating as input to control the followings algorithms:

1) OpenH323: an open source and packet loss intolerant algorithm, used in CallGen323 application;
2) Window: histogram based algorithm with spike detection, presented in [3];
3) Adaptive: algorithm proposed by [14], which is also reactive and quality based, that tries to maximize the end-user perceiving quality.

III. THEORETICAL ASPECTS OF BUFFER DELAY

In the next Sections, consider n^k the set of packets belonging to k-th talkspurt and p_i^k, a_i^k and t_i^k, respectively, the playout, receiver and transmission time. Using de-jitter buffer (or buffer delay - BD) in receiver side, with dynamic adjustment to each talkspurt, the playout time of i-th packet is:

$$p_i^k = a_1^k + BD^k + (i-1)\Delta t_i^k \quad (3)$$

where $\Delta t_i^k = t_i^k - t_{(i-1)}^k$.

 ISBN: 978-1-61208-411-4

A packet will be lost when it does not meet the jitter restriction [15] [16], i.e., BD is not enough to jitter removal in packet i, then:

$$p_i^k > a_1^k + BD^k + (i-1)\Delta t_i^k \quad (4)$$

Theorem 1 presents a buffer delay value to prevent packet loss by jitter.

Theorem 1: In a talkspurt k, with buffer delay BD^k, no packet is lost by jitter restriction violation if and only if

$$BD^k \quad \geq \max_{i\in\{1,2,...,n^k\}}\{\delta_i^k - (i-1)\Delta t_i^k\}$$

where $\delta_i^k = a_i^k - a_1^k$ for every $i \in \{1,2,...,n^k\}$.

Proof: Since there is no packets loss in talkspurt, this is equivalent to: $p_i^k - a_i^k \geq 0$ for every $i \in \{1,2,...,n^k\} \Leftrightarrow a_i^k \leq p_i^k \Leftrightarrow a_i^k \leq a_1^k + BD^k + (i-1)\Delta t_i^k \Leftrightarrow a_i^k - a_1^k \leq BD^k + (i-1)\Delta t_i^k \Leftrightarrow BD^k \geq (a_i^k - a_1^k) - (i-1)\Delta t_i^k \Leftrightarrow BD^k \geq \max_{i\in\{1,2,...,n^k\}}\{\delta_i^k - (i-1)\Delta t_i^k\}$, for every $i \in \{1,2,...,n^k\}$.

Notice that:

$$BD^k \geq BD_{npl}^k = \max_{i\in\{1,2,...,n^k\}}\{\delta_i^k - (i-1)\Delta t_i^k\}$$

where BD_{npl}^k is the buffer delay which does not present packet loss. ■

Thus, we introduce the notion of limiting due to jitter. In the next definitions consider $N = \{1,2,...,n^k\}$ all packet indexes of the k-th talkspurt.

Definition 1: The BD_c^k is c-th limiting due to jitter, i.e., the value that remove jitter in a set Ω_c of packets of talkspurt k is defined by

$$BD_c^k = \max_{i\in\Omega_c}\{\delta_i^k - (i-1)\Delta t_i^k\},$$

where $\Omega_0 = N$, and $\Omega_c = N - (u_0 \cup u_1 \cup ... \cup u_{c-2} \cup u_{c-1})$ for $c > 0$, and $u_c = \{r_c^1, ..., r_c^{w_c}\}$ are the w_c packets where $p_i^k = a_1^k + BD_c^k + (i-1)\Delta t_i^k$ with $i \in u_c$.

Lemma 1: There is a finite number of jitter limiting values in a talkspurt.

Proof: The first jitter limiting value is: $BD_0^k = \max_{\{i\in\Omega_0=N\}}\{\delta_i^k - (i-1)\Delta t_i^k\}$, used by set of packets $u_0 \subset \Omega_0 = N$. Consider $\Omega_1 = N - u_0 \subseteq \Omega_0$, if $\Omega_1 = \emptyset$, the proof is completed, otherwise it is possible to calculate other jitter limiting value: $BD_1^k = \max_{\{i\in\Omega_1\}}\{\delta_i^k - (i-1)\Delta t_i^k\}$ for which there is a non-empty set $u_1 \subseteq \Omega_1 \subset \Omega_0$ of packets. This reasoning is applied until one is found $\Omega_{m+1} = \emptyset$, then the last jitter limiting value is $BD_m^k = \max_{\{i\in\Omega_m\}}\{\delta_i^k - (i-1)\Delta t_i^k\}$, where $m \leq n$ and, exist $\emptyset \neq u_m \subseteq \Omega_m \subset \Omega_{m-1} \subset ... \subset \Omega_0$ of packets that use that value to remove jitter. Thus, we obtain a finite number of jitter limiting value.

■

Lemma 2: The jitter limiting values are presented in the format $BD_j^k < BD_{j-1}^k$ to $j = 1, ..., m$.

Proof: Considering $BD_j^k < BD_{j-1}^k$, then:

$$BD_j^k = \max_{\{i\in\Omega_j\}}\left\{\delta_i^k - (i-1)\Delta t_i^k\right\}$$

$$BD_{j-1}^k = \max_{\{i\in\Omega_{j-1}\}}\left\{\delta_i^k - (i-1)\Delta t_i^k\right\}$$

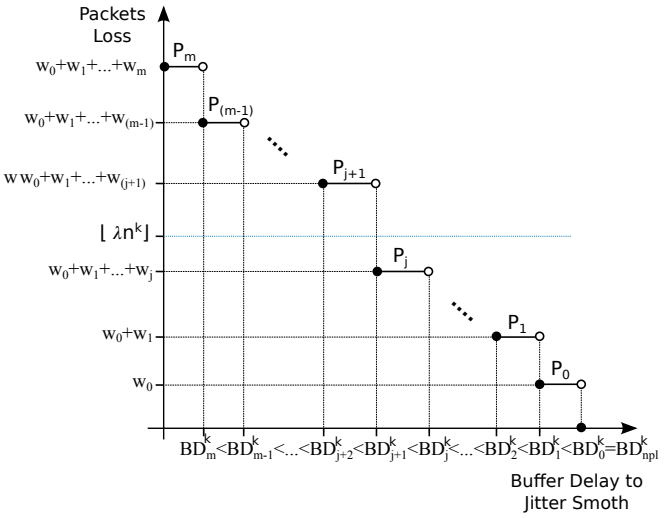


Figure 2. Steps due to jitter in a talkspurt.

where $\Omega_j = N - (u_0 \cup ... \cup u_{j-2} \cup u_{j-1})$ and $\Omega_{j-1} = N - (u_0 \subset \cup ... \cup u_{j-2})$, so $\Omega_j \subset \Omega_{j-1}$, then $BD_j^k < BD_{j-1}^k$. If $BD_j^k = BD_{j-1}^k$ then $u_j \cap < u_{j-1} \neq \emptyset$ ■

In a talkspurt, we have the following ordering $BD_m^k < BD_{m-1}^k < ... < BD_0^k$.

Definition 2: Intervals of type $P_m = \left[0, BD_m^k\right), P_{m-1} = \left[BD_m^k, BD_{m-1}^k\right), ..., P_0 = \left[BD_1^k, BD_0^k\right)$ will be referenced as steps due to jitter.

Definition 3: At each step, due to jitter we have associated a number named degree, given by:

$$degree(P_j) = \sum_{c=0}^{j} w_c$$

where $BD_{m+1}^k = 0$, and $j = 0, ..., m$.

Each degree is unique by definition. Besides $w_c \geq 1$ for each c, from what we can conclude that:

$$0 < degree(P_0) < ... < degree(P_m)$$

The lemma 3 allows monitoring the packet loss behaviour with each BD^k value used to jitter remove.

Lemma 3: Using BD^k in a talkspurt, then the number of lost packets is equal to the degree to which the step belongs.

Proof: In a talkspurt, we have the degrees P_j, with $j = 0, ..., m$, due to lemma 2 we have that $0 = BD_{m+1}^k < BD_m^k < ... < BD_{j+1}^k \leq BD^k < BD_j^k < ... < BD_1^k < BD_0^k$. Then, $\max_{\{i\in\Omega_{j+1}\}}\left\{\delta_i^k - (i-1)\Delta t_i^k\right\} = BD_{j+1}^k \leq BD^k < BD_j^k < ... < BD_0^k$ and:

$$\begin{aligned} BD_j^k &= \delta_r^k - (r-1)\Delta t_r^k,\ r \in u_j \\ BD_{j-1}^k &= \delta_r^k - (r-1)\Delta t_r^k,\ r \in u_{j-1} \\ &\vdots \\ BD_0^k &= \delta_r^k - (r-1)\Delta t_r^k,\ r \in u_0 \end{aligned}$$

Assuming that $BD^k < (a_r - a_1) - (r-1)\Delta t_r^k$ for all $r \in u_0 \cup ... \cup u_j$, thus we have, $a_1 + BD^k + (r-1)\Delta t_r^k < a_r$, for $r \in u_0 \cup ... \cup u_j$, i.e., the jitter restriction is broken for all $r \in u_0 \cup ... \cup u_j$, then packets $r_j, ..., r_0$ are lost. On the

other hand, with $BD^k \geq (a_r - a_1) - (r-1)\Delta t_r^k$, for all $r \in \Omega_{j+1} \supset \Omega_{j+2} \supset ... \supset \Omega_m \supset \Omega_{m+1}$.Then $p_r \geq a_r$ for all $r \in \Omega_j$, and $r \in \Omega_{j+2}$, so on for all $r \in \Omega_m$. With $u_{j+1} \subseteq \Omega_{j+1}, ..., u_m \subseteq \Omega_m$, the packets $r_{j+1}, ..., r_m$ not be lost, and $\{u_0, ..., u_m\}$ a subset of *N*, the total number of packets is $w_0 + ... + w_j = \sum_{c=0}^{j} w_c = degree(P_j)$. ■

IV. Buffer Delay Correction Algorithm

Prior to present the BDCA, an important definition is presented that relates buffer delay and target packet loss. This value is named OBD. In the previous Section, we can see that there is a certain limit to Buffer Delay (BD^k), and above this level there is no packet loss. On the other hand, the good quality of voice communication admits a certain limit of packet loss. Therefore, let us suppose a $\lambda \in (0, 1)$ of packets loss in a talkspurt, i.e., at most $\lfloor \lambda n^k \rfloor$ packets can be lost (see Figure 2) where $\lfloor x \rfloor$ is the floor function (greater integer smaller than or equal to x). In this case we are interested in solving (5) bellow.

$$\min \left\{ f(BD^k) = BD^k \mid \Psi(BD^k) \leq \lfloor \lambda n^k \rfloor, BD^k \in [0, +\infty) \right\} \quad (5)$$

That is the optimum buffer delay (OBD_λ^k), which represents a minimum delay value inserted in a talkspurt k, with target loss factor λ.

Theorem 2: In a talkspurt that use BD^k, n' packets will be lost, if and only if, BD^k belongs to degree of with step n'.

Proof: When $n' = 0$, i.e., no packet is lost, the theorem 1 assure this proof. If $n' > 0$, consider $W = \{w_0, w_0 + w_1, ..., w_0 + ... + w_m\}$ a set of all packets lost by jitter, if $n' \in W$ with $n' = w_0 + w_1 + ... + w_j$ for any j, then $BD^k \in P_j$. If $BD^k \in P_h$ for $0 \leq h < j$, less than n' packets would be lost, on other hand, if $j \leq h < m$, more than n' would be lost. With $degree(P_j) = w_0 + ... + w_j$, the BD^k belongs to a degree, with step n'. ■

Looking for theorem 2 and $BD^k \subset \{P_m, ..., P_0\}$ with $P_i \in [0, +\infty)$ we can write (5) as follow:

$$\min \left\{ \min\{ f(BD^k) = BD^k | \Psi(BD^k) \leq \lfloor \lambda n^k \rfloor, BD^k \in I \} \right\} \quad (6)$$

where $I \in \{P_m, ..., P_0\}$ and $\min\{f(BD^k) = BD^k \mid \Psi(BD^k) \leq \lfloor \lambda n^k \rfloor, BD^k \in I\}$ can be solved by Weierstrass Theorem, because in this case, I is compact and f is continuous.

The BDCA is a method to adjust the value presented by one BDA, i.e., approaching BD_{BDA}^k to OBD_λ^k, with packet loss rate in λ. To apply BDCA over talkspurt k, the Adjust Factor (AF) is computed as:

$$AF(k) = \frac{1}{Z} * \sum_{i=(k-1-Z)}^{i=(k-1)} \frac{OBD_\lambda^i}{BD_{BDA}^i} \quad (7)$$

The window size (Z) has the last 40 received talkspurts to reduce computational costs, values greater than 40 do not change significantly the results. To find OBD_λ^i, the following elements are needed:

- Packets transmitted until talkspurt $(i-1)$;

$$N_{i-2} = \sum_{j=1}^{i-2} n^j \quad (8)$$

- Number of packets lost from talkspurts 1 to $(i-1)$;
- Target Packet loss rate (λ).

The OBD_λ^i should be used in i-th talkspurt to bring the packet loss closer to λ. The BD_{BDA}^i is the value computed by selected BDA. Equation (9) shows Buffer Delay adjusted:

$$BD_{BDCA}^k = BD_{BDA}^k * AF(k) \quad (9)$$

Then, the adjusted playout time ($\hat{p}_i^k$) is defined by:

$$\hat{p}_i^k = a_1^k + BD_{BDCA}^k + (i-1)(\Delta t_i) \quad (10)$$

Consider the talkspurt $(k-1)$ received, the BDCA to playout time adjust of talkspurt k is showed in Figure 3.

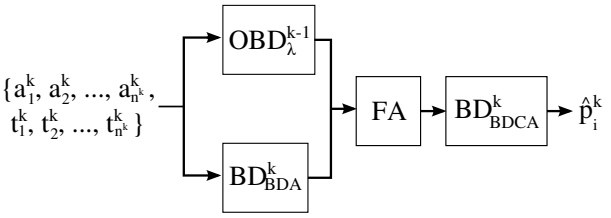


Figure 3. BDCA.

The OBD algorithm presents linear time and can run in parallel with BDCA, this makes BDCA defined by BDA's computational complexity.

V. Implementation and Results

In this Section, a performance analysis of the proposed algorithm is presented. The BDAs used for comparison are:

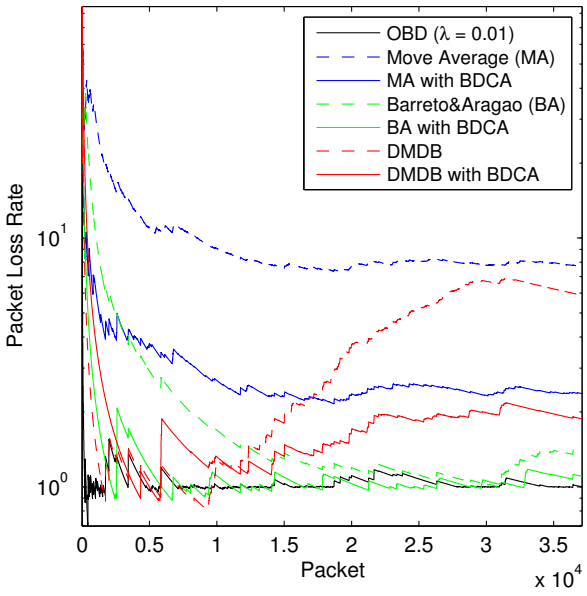
- Move Average Algorithm (MA) [10] a loss-tolerant technique;
- Algorithm from Barreto and Aragao (BA) presented in [8], classified as loss-intolerant technique;
- Dynamic Management of Dejitter Buffer (DMDB) presented in [13], considered a quality based technique.

For the tests, we consider the traces described in [3]. The traces contain the sender and receiver timestamps of transmitted packets. One 160 bytes audio packet is generated approximately at every 20 ms when there is speech activity [17]. The number of concurrent applications, network protocols or other elements of network environment may change the packet delay, but do not affect the BDCA. This enable the use of traces in simulation tests. A description of the traces is depicted in Table I.

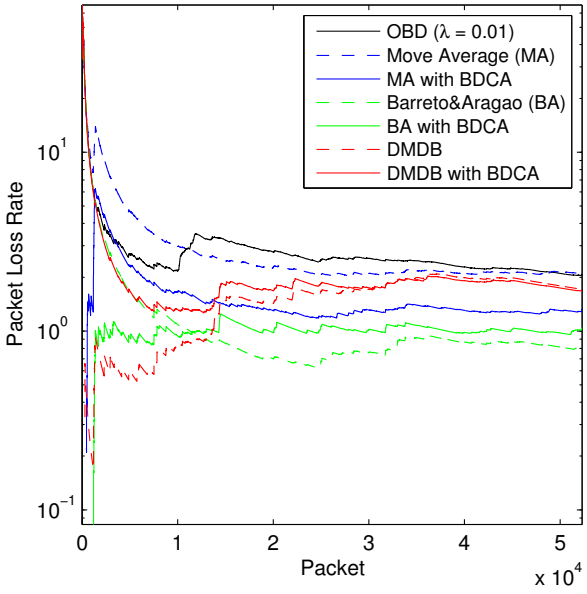
TABLE I. Traces Description.

trace	*Talkspurts*	*Packets*	*Length (s)*
A	536	37104	165.696
B	540	52296	174.604

To assess the performance of BDCA, we focus in packet loss rate, buffer delay average and quality of call (MOS). Considering N packets in a session, M the number of talkspurts, n^k the number of packets in talkspurt k, r_i^k the success indicator with values $r_i^k = 0$ when the packet is lost ($p_i^k < a_i^k$) or $r_i^k = 1$ when packet is available in receiver on playout time



(a) Trace A



(b) Trace B

Figure 4. Traces A and B with $\lambda = 0.01$

$(p_i^k \geq a_i^k)$. The total number of packets played out in an audio session is given by

$$\Upsilon = \sum_{k=1}^{M} \sum_{i=1}^{n^k} r_i^k \tag{11}$$

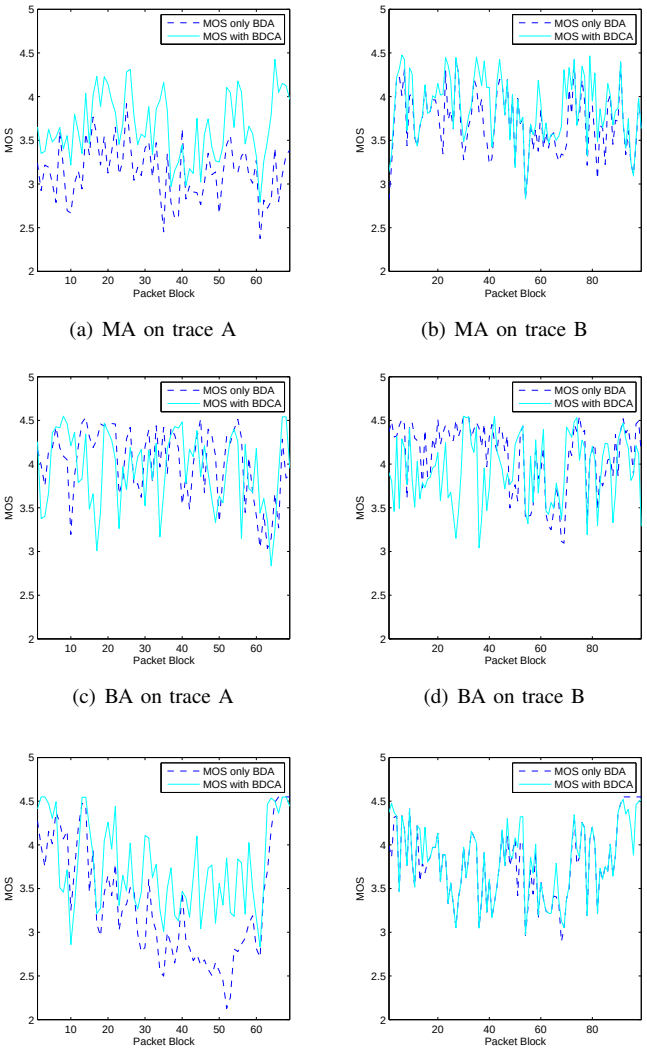
We consider in this work the average buffer delay to remove jitter (BD_{av}), shown in (12).

$$BD_{av} = \frac{1}{\Upsilon} \sum_{k=1}^{M} \sum_{i=1}^{n^k} r_i^k \left(p_i^k - a_i^k\right) \tag{12}$$

The percentage of packets not used in audio application on the receiver side (ω) is obtained by the (13).

$$\omega = \frac{N - \Upsilon}{N} * 100 \tag{13}$$

In graphs of Figure 4 we use the terms "With BDCA" to represent the original BDA running with BDCA. The target percentage of packets loss is 1%. These graphs are showing the evolution of packet loss rate in a voice call. For interactive audio, packet loss rate is considered adequate up to 1% of call [18], [19]. The Figure 5 are showing MOS values computed using PESQ algorithm over selected BDAs.



(a) MA on trace A (b) MA on trace B

(c) BA on trace A (d) BA on trace B

(e) DMDB on trace A (f) DMDB on trace B

Figure 5. Evolution of MOS.

TABLE II. PACKET LOSS TARGET (λ) IN 1%.

trace	BDA	Only BDA			BDA with BDCA			OBD		
		ω	BD_{av}	MOS	ω	BD_{av}	MOS	ω	BD_{av}	MOS
A	MA	7.70	116.55	3.13	2.37	180.52	3.65	0.99	121.74	4.01
	BA	1.39	216.59	4.01	1.10	401.93	3.93	0.99	121.74	4.01
	DMDB	5.88	191.97	3.34	1.87	290.62	3.75	0.99	121.74	4.01
B	MA	2.13	42.45	3.69	1.31	61.62	3.87	2.08	30.78	3.74
	BA	0.83	65.94	4.10	1.03	68.99	3.93	2.08	30.78	3.74
	DMDB	1.72	64.22	3.78	1.70	79.81	3.82	2.08	30.78	3.74

 ISBN: 978-1-61208-411-4

The Table II shows the results of packet loss target percentage (1%), the columns ω and BD_{av} are expressed in percentage of transmitted packets and milliseconds, respectively. The MOS column present an average value computed using PESQ algorithm [20], on blocks of 3000 packets, with shift of 500 packets to next window. The tests were made in Matlab [21].

VI. Conclusion

In this paper, we presented the Buffer Delay Correction Algorithm (BDCA) to reduce the difference between packet loss rate of any BDA and the Optimum Buffer Delay (OBD). We have compared the BDA with and without BDCA using 1% of packet loss rate.

Figures 4(a) and 4(b) show that packet loss percentage (ω) with BDCA are closer to values from OBD than running only BDA. But any greater buffer delay is able to produce a reduced packet loss rate. The BDCA uses only the necessary buffer delay to regulate the packet loss rate to closer to target value. This can be viewed in Table II and graphics of Figure 5, call quality is best or equal the results "without BDCA" (or only BDA) in most parts of calls.

We are currently expanding the definitions of Buffer Delay Adjustment to reach packet loss caused by latency, i.e., including the sum of packet discarded with playout time greater than the maximum threshold (L). To reach this new restriction, we are working in a new formulation of Adjust Factor.

References

[1] W. C. Hardy, VoIP Service Quality: Measuring and Evaluating Packet-Switched Voice). McGraw-Hill, 2003.

[2] Z. Qiao, R. K. Venkatasubramanian, L. Sun, and E. C. Ifeachor, "A new buffer algorithm for speech quality improvement in voip systems," Wirel. Pers. Commun., vol. 45, no. 2, apr 2008, pp. 189–207. [Online]. Available: http://dx.doi.org/10.1007/s11277-007-9408-7

[3] S. B. Moon, J. Kurose, and D.Towsley, "Packet audio playout adjustment: performance bounds and algorithms," ACM/Springer Multimedia Systems, vol. 6, no. 1, january 1998, pp. 17–28. [Online]. Available: http://dx.doi.org/10.1007/s005300050073

[4] Y. Zhang, D. Fay, L. Kilmartin, and A. W. Moore, "A garch-based adaptive playout delay algorithm for voip," Computer Networks: The International Journal of Computer and Telecommunications Networking, vol. 54, no. 17, dec 2010, pp. 3108–3122. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2010.06.006

[5] L. Atzori and M. L. Lobina, "Playout buffering in ip telephony: a survey discussing problems and approaches," Communications Surveys Tutorials, IEEE, vol. 8, no. 3, 2006, pp. 36–46. [Online]. Available: http://dx.doi.org/10.1109/COMST.2006.253269

[6] C. Perkins, RTP: Audio and Video For The Internet. Addison Wesley, 2012.

[7] R. Ramjee, J. Kurose, D. Towsley, and H. Shulzrinne, "Adaptive Playout Mechanisms for Packetized Audio Applications in Wide-Area Networks," in Proceedings of IEEE Infocom, vol. 2, Montreal, Canada, 1994, pp. 680–688. [Online]. Available: http://citeseer.nj.nec.com/ramjee94adaptive.html

[8] J. B. A. Jr. and G. A. Barreto, "Novel approaches for online playout delay prediction in voip applications using time series models," Computers and Electrical Engineering, vol. 36, no. 3, 2010, pp. 536 – 544. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S004579060900113X

[9] K. Fujimoto, S. Ata, and M. Murata, "Statistical analysis of packet delays in the internet and its application to playout control for streaming applications," IEICE Transactions on Communications, vol. E84-B, no. 6, june 2001, pp. 1504–1512. [Online]. Available: http://www-ana.ist.osaka-u.ac.jp/achievements/web2001/papers/k-fujimo01ieice-ModelingPlayout.pdf

[10] V. M. R. Ramos, C. Barakat, and E. Altman, "A moving average predictor for playout control in voip," in Proceedings of the 11th international conference on Quality of service, ser. IWQoS'03, vol. 2707. Berlin Heidelberg: Springer-Verlag, June 2003, pp. 155–173. [Online]. Available: http://dl.acm.org/citation.cfm?id=1784037.1784049

[11] K. Fujimoto, S. Ata, and M. Murata, "Adaptive playout buffer algorithm for enhancing perceived quality of streaming applications," Telecommunication Systems, vol. 25, no. 3-4, april 2004, pp. 259–271.

[12] ITU-T P.800, Recommendation ITU-T P.800 - Methods for subjective determination of transmission quality, Telecommunication Standardization Sector of International Telecommunication Union (ITU), august 1996. [Online]. Available: http://www.itu.int/rec/T-REC-P.800

[13] R. F. Valle, L. S. G. de Carvalho, R. B. Aguiar, E. S. Mota, and D. Freitas, "Dynamical management of dejitter buffers based on speech quality," in Proceedings of the The IEEE symposium on Computers and Communications, ser. ISCC '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 56–61. [Online]. Available: http://dx.doi.org/10.1109/ISCC.2010.5546799

[14] L. Sun and E. C. Ifeachor, "Prediction of perceived conversational speech quality and effects of playout buffer algorithms," in Communications, 2003. ICC '03. IEEE International Conference on, vol. 1, 2003, pp. 1–6.

[15] F. Sakuray, R. S. V. Hoto, and L. S. Mendes, "Analysis and estimation of playout delay in voip communications," International Journal of Computer Science and Network Security, vol. 8, no. 3, March 2008, pp. 98–105. [Online]. Available: http://paper.ijcsns.org/07_book/200803/20080315.pdf

[16] D. Florencio and L.-W. He, "Enhanced adaptive playout scheduling and loss concealment techniques for voice over ip networks," in Circuits and Systems (ISCAS), 2011 IEEE International Symposium on, 2011, pp. 129–132. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5937518

[17] H. Schulzrinne, "Voice communication across the internet: a network voice terminal," Tech. Rep., july 1992. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.120.1343

[18] S. Nagireddi, VoIP Voice and Fax Signal Processing, 1st ed. Wiley Publishing, 2008.

[19] TIA/EIA 116A, Telecommunications-IP Telephony Equipment - Voice Quality Recommendation for IP Telephony, Telecommunication Industry Association, 2006.

[20] ITU-T P.862, Recommendation ITU-T P.862 - Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speechquality assessment of narrow-band telephone networks and speech codecs, Telecommunication Standardization Sector of International Telecommunication Union (ITU), february 2001. [Online]. Available: http://www.itu.int/rec/T-REC-P.862

[21] MATLAB, MATLAB and Statistics Toolbox Release 2012b. Natick, Massachusetts: The MathWorks Inc., 2012.

 ISBN: 978-1-61208-411-4

Using Firefly and Genetic Metaheuristics for Anomaly Detection based on Network Flows

Fadir Salmen, Paulo R. Galego Hernandes Jr.

Security Information Department
São Paulo State Technological College (FATEC)
Ourinhos, Brazil
Email: {fadirsalmen, paulogalego}@gmail.com

Luiz F. Carvalho, Mario Lemes Proença Jr.

Computer Science Department
State University of Londrina (UEL)
Londrina, Brazil
Email: luizfcarvalhoo@gmail.com,proenca@uel.br

***Abstract*—Traffic monitoring is a challenging task which requires efficient ways to detect every deviation from the normal behavior on computer networks. In this paper, we present two models to detect network anomaly using flow data such as bits and packets per second based on: Firefly Algorithm and Genetic Algorithm. Both results were evaluated to measure their ability to detect network anomalies, and results were then compared. We experienced good results using data collected at the backbone of a university.**

Keywords-Anomaly Detection; Traffic Monitoring; Network Management; Genetic Algorithm, Firefly Algorithm.

I. INTRODUCTION

Managing a network is a complex job and requires support from a number of tools and techniques, which help manage the resources efficiently. Administrators must have a smart use of bandwidth resources, identifying anomalous traffic without human supervision.

A Denial of Service (DoS) attack can be the reason for an unavailable network. The objective of a DoS is to crash a service by attempting to reach the machine's access limit. An attacker sends packets labeled to specific IP and port addresses, simulating a legitimate access, but it sends a huge quantity of packets, with the only intent of bring down a server or service, making it impossible for a real person to access this service. A Distributed Denial of Service (DDoS) attack uses multiple compromised systems to launch several DoS attacks, coordinated against one or more victims. In fact, a DDoS attack adds the many-to-one dimension to the DoS problem [1].

For many years, network administrators used to get their technical information using the Simple Network Management Protocol (SNMP). However, this protocol could not present many details about the real network usage due to its limited set of features. With the use of data flow, administrators could obtain more knowledge about their environments. A flow record is defined by a connection between two peers reporting fields in common, those could be the endpoint addresses, protocol, time, and volume of information transferred. This gives a more detailed view on the traffic and permits using it on large networks, due to the data reduction compared to SNMP [2].

In order to identify an anomaly, we have to know what is considered normal behaviour in the network. When the normal behavior is described, every deviation of this profile can be virtually described as an anomaly. A network anomaly detection system has to work without any supervision, and have to avoid security incidents, being useful and effective in order to keep the network available as frequently as possible.

There are some tools used by network managers to identify attacks in their environments. According to Teodoro *et al.* [3] there are signature-based systems, whose detection process is generally fast and reliable because of the usual pattern-matching procedure considered in the detection stage. Nevertheless, the signature database has to be updated every moment and a signature-based system is unable to detect attacks previously unobserved.

To overcome this lack of security, there are models based on traffic characterization, which are able to learn from the normal behavior of an environment, and based on its history, detect every change in the network routine. In this paper, we present a model to identify anomalous network traffic, based on traffic characterization, which uses the Firefly Algorithm (FA) to classify network flows, and compare this model with another method, based on Genetic Algorithm (GA). Our goal is to create a Digital Signature of Network Segment using Flow Analysis (DSNSF) utilizing both GA and FA, and use this DSNSF to identify anomalous traffic through the creation of a threshold. We use a real set of data to perform the process and evaluate the results to prove the accuracy of our DSNSF models. Also, we compared these two methods to identify the advantages and disadvantages of each one.

The metaheuristics FA and GA have powerful and distinct techniques in the optimization of an objective function, specially for a wide search space. Thus, a comparative study of these algorithms, measuring their efficiency and quality to detect anomalies in computer networks was necessary.

This paper is organized as follows: Section II presents the related work. Section III explains the DSNSF-GA method giving details of the DSNSF-FA generation. Section IV discusses the result of our evaluation tests, and finally Section V concludes this paper.

II. RELATED WORK

FA is an algorithm based on the fireflies behavior and its emitted light characteristics. In the study presented by Gandomi *et al.* [4], they used Firefly Algorithm (FA) to

 ISBN: 978-1-61208-411-4

efficiently solve several variable issues to structural engineering optimization. Despite its restrictions, FA was used in order to decrease the following production cost: physical characteristics of beams, cylindrical pressure vessel, helical compression spring design and a reinforced concrete beam design, besides helping the development of an automotive side impact protection.

In their study, Hassanzadeh *et al.* [5] used FA algorithms, due to its high convergence features with low processing time, to optimize Otsu's method on image segmentation. Research results showed the efficiency and accuracy of the method for segmentation.

The GA is an evolutionary algorithm developed by Holland [6], which is based on the natural evolution of species. Based on operators such as selection, crossover and mutation, GA is recognized as an ideal optimization technique to solve a large variety of problems, such as organizing data under some condition or optimizing search problems. In [7], a genetic algorithm was used to organize data in clusters, when the task of GA was to search for the appropriate cluster centers.

An anomaly detection system was proposed in [8], which utilizes the SNMP protocol and searches for a correlation on the behavior of some SNMP objects, avoiding the high rate of false alarms. Another work using correlation was found in [9], which utilizes the observation among the network nodes, measuring delays and drop rates between each connection. To characterize network traffic, certain techniques could be applied such as Holt-Winters for Digital Signature, a modification of the classic statistical method of forecasting Holt-Winters [10]. In [11], the Autoregressive Integrated Moving Average (ARIMA) was used to generate forecasts for data segments. The author introduces the use of a non-classical logic called Paraconsistent Logic to improve the DSNSF employment.

III. The Generation Of DSNSF

The target of our work is to permit network administrators to identify anomalous behavior in their environments based on traffic characterization. For this purpose, we created a DSNSF, which was introduced by Proença et al. [12] in which a Digital Signature of Network Segment (DSNS) was generated using historical traffic of workdays to describe the normal network usage for subsequent weeks. Research extended and improved by [13] and [14].

In this paper, we present two metaheuristic strategies to create a DSNSF using data as bits and packets. These data were collected from the networks assets using sFlow, a standard for monitoring high-speed switched and routed networks, which uses the sampling technique to collect flows [15]. Our purpose in this work is to demonstrate that these two flow attributes, bits and packets per second can be used to identify a normal, or expected, traffic pattern and consequently appoint every network anomaly in the traffic. The first model is based on fireflies behavior and its emitted light characteristics, and is used to optimize the K-means clustering algorithm. The second model is based on the natural evolution of species theory, implemented in computing as Genetic Algorithm, which simulates the natural process of evolution in a population. Both methods are appropriate to the DSNSF construction and they will be described ahead.

A. DSNSF-FA

DSNSF-FA is an algorithm developed to construct a normal network behavior profile, based on the network traffic patterns recognition and that will enable the creation of an anomaly detection system.

The DSNSF-FA structure is based on two other algorithms, k-means, used to clustering and FA, on the determination of centroids, which will be the points responsible for the construction of DSNSF. A centroid is a point which indicates the center of the cluster. This combination is required, due to a shortcoming presented by k-means, which is solved by FA. According to Gungor and Unler [16], k-means presents a big problem in its algorithm, which is related with the centers startup. If the centers are started very close, k-means will converge to a minimum local.

1) Firefly Algorithm: The optimization process is present in every system where you want to achieve certain goals, being on the professional range, searching a lower production cost or even in vacation planning, determining the shortest path to the desired place. Before several algorithms, the use and application of metaheuristic algorithms based on nature has grown, among them is the Firefly Algorithm (FA) [17].

The optimization performed by the algorithm FA is based on the attraction between fireflies. The lower brightness firefly will position even closer to a firefly with higher luminescence and when it does not find a brighter firefly, it will randomly move until it finds a brightness that attracts it. This behavior will repeat until every firefly gets together and then this place become the best solution, in other words, optimize an objective function [18].

2) K-means: K-means is an unsupervised clustering method, whose function is to group similar items in subgroups (clusters). Thus, this enables the partitioning R records into K groups, being $R > K$, where the distance between all the resulting data of a subgroup and its said center, summed by all subgroups, to be minimized.

An easy implementation and high-speed K-means was proposed by Macqueen [19], in which objective function is shown by Equation 1:

$$KM_{(x,c)} = \sum_{i=1}^{n} \sum_{j=1}^{k} |x_i - c_j|^2 \quad (1)$$

Where x is the data vector and c is the vector of centers, n is the number of elements on x and k is the number of centers on c.

3) DSNSF-FA model: DSNSF-FA works with historical database, arranged in time frames of 5 minutes. We found in previous works [10] [20–22], that 5 minutes is an ideal interval, however using sFlow we are dealing with sampling of data. A 5 minutes interval, preserves the exportation pattern used by Nfdump [23]. For each workday in a week, we gathered data from their equivalent counterparts in the three previous weeks. That is, if a Monday is analyzed, the historical database to be used will be related to the previous three Mondays. This database will be divided into three clusters, according to similarities defined by K-means. For each one of the clusters, FA will determine its best representative, in other words, the centroid. This operation is performed with the optimization

 ISBN: 978-1-61208-411-4

of the chosen objective function. The DSNSF-FA works as objective function such as the Euclidean Distance, presented by the Equation 2:

$$D_{ij} = \sum_{i=1}^{Q} \sum_{j=1}^{K} \sqrt{\sum_{n=1}^{d} (x_{in} - c_{jn})^2} \quad (2)$$

in which Q is the amount of data to be clustered, K is the total of clusters, d the dimension, x_in indicates the data value i on n and c_jn is clusters center value j on dimension n. At the end of the iterations, there will be three centroids, one to each cluster defined by K-means. For each one of these centroids, the DSNSF-FA will assign a weight to theirs luminosities, defined by Equation 3:

$$Lic_k = Lrc_k * (nc_k/N) \quad (3)$$

according to the amount of data each one represents, in which Lrc_k corresponds to the resident brightness of the cluster centroid k, N to the total amount of fireflies by iteration and nc_k refers to the amount of fireflies of cluster k, and then FA is applied on these three centers, resulting in the representative centroid of the data initially selected. Therefore, the first point of DSNSF will be generated. This approach will be held until the entire historical database is processed and the points which will generate the DSNSF are known.

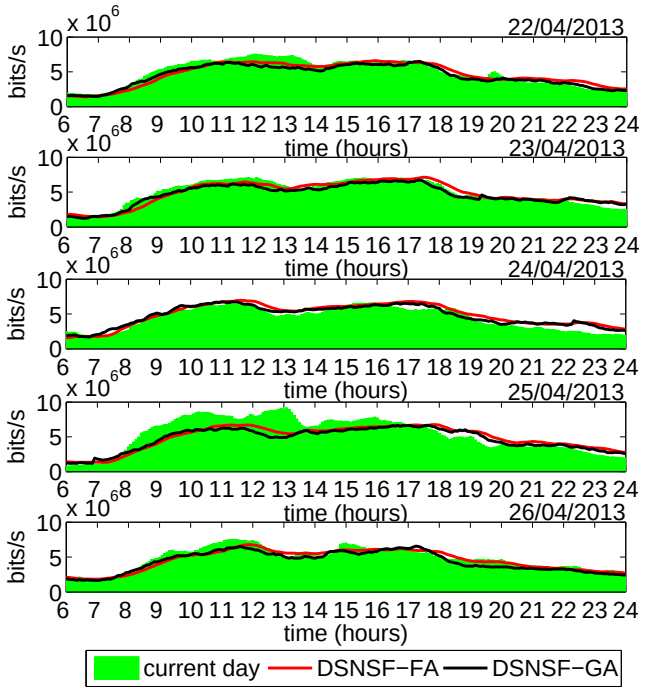


Figure 1. DSNSFs for bits/s - from 22^{nd} to 26^{th} April, 2013.

For the creation of DSNSF-FA, we used IP flows of historical data of State University of Londrina (UEL). These data were collected and stored in a historical basis for future reference and when requested, are delivered in files. The files were used containing bits and packets quantities, collected per second, using workdays from 22^{nd} April to 3^{rd} May of 2013, which served as the learning process and creation of DSNSF-FA. The DSNSF-FA, then, was superimposed on the real traffic, where it was possible to observe the traffic network anomalies.

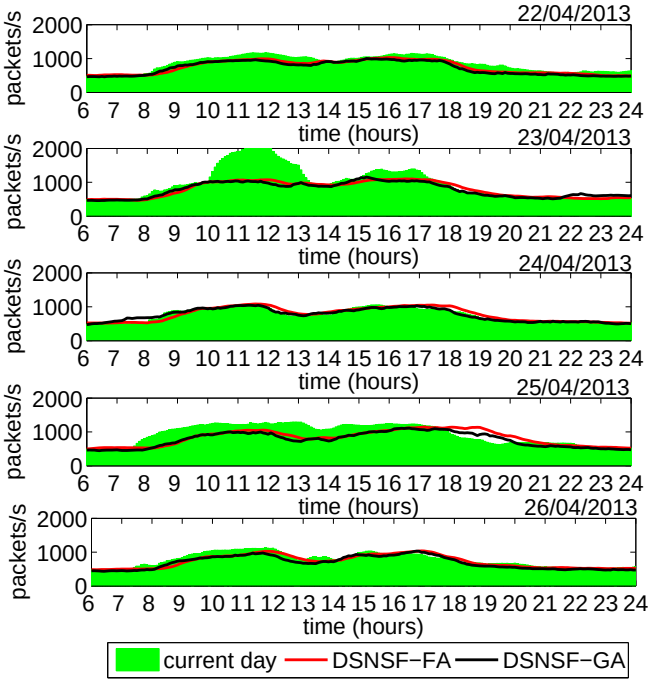


Figure 2. DSNSFs for packets/s - from 22^{nd} to 26^{th} April, 2013.

The DSNSF-FA algorithm operation is shown by DSNSF-FA Algorithm (1).

Algorithm 1 – DSNSF-FA

Require: set of bits and packets collected from historical database

Ensure: X: Vector representing the normal behavior for bits and packet sets of a day arranged in 288 intervals of 5 minute, i.e. the DSNSF

1: **for** $i = 1$ to 288 **do**
2: Applies K-means, K=3
3: **for** $t = 1$ to number of iterations **do**
4: Applies FA for each cluster
5: Calculate the center of each cluster of the best solution - objective function
6: **end for**
7: For each center, applies weight function
8: **for** $t = 1$ to number of iterations **do**
9: Applies FA to the three centers, K=1
10: Calculate the center of cluster of the best solution - objective function
11: **end for**
12: $X_i \leftarrow$ Average among the clusters
13: **end for**
14: **return** X

Initially, the information contained in the files are prepared to provide data every 5 minutes, generating 288 samples. These data are initially processed by K-means algorithm, which distributes them in three clusters. The $K = 3$ choice was the result of the interpretation and validation of cluster, for

 ISBN: 978-1-61208-411-4

the amount of data to be analyzed, performed by methods of Silhouette, Davies Bouldin, Calinski Harabasz, Dunn and Krzanwki Lai [24].

In each cluster, the FA algorithm is applied to find its respective centroid. This process optimizes the objective function used, where the luminosity of fireflies relates directly. After obtaining the three centroids, a weight is assigned to each one according to the amount of data they represent on their residual luminosity.

Then, the FA algorithm is used on the three centroids in order to find the result of the first 5 minutes sample analyzed. This centroid is responsible for the first data point of DNSNF-FA. In sequence, it will start the analysis of the other 287 samples, arriving at a total of 288 data points, which will then allow for the construction of the desired DSNSF-FA.

B. DSNSF-GA

The DSNSF-GA, presented in [22] uses a genetic algorithm based approach to organize data flow in clusters. Each cluster has its own centroid, and we measure the distance between the points to organize data and use the average among centroids to generate our DSNSF. The rule was the same for the DSNSF-FA, so for each workday in a week, we used data from the same day in the last three weeks, and compare them with the current day.

GA manipulates a population of potential problem solutions, trying to solve them using a coded representation of these solutions, which is the equivalent to genetic material (chromosomes) of individuals in nature. In GA, members of a population (the solutions) compete with each other to survive, reproduce and generate new solutions, using operators such as selection, crossover and mutation.

To start the process, we generate a random initial population in which we began applying the three operators. Our chromosomes have cluster centroids values. We appointed an initial population of forty parents. They create the new generation, which will replace the old one. It will repeat for a fixed number of iterations. At the end of this process, we have the best chromosomes based on their fitness function, which is the Euclidean Distance, the same as the FA algorithm. This value represents a single point in the DSNSF-GA. We have to apply the clusterization using GA for each point in the graphic, so it will be repeated for 288 times, one point every five minutes. Using the Silhouette method for interpretation and validation of clusters, best results were reached using $K = 3$.

To yield new generations, the crossover operator will combine chromosomes of two parents to create a new one. This process will continue until the old population be replaced by a new population of children. As in nature, the fittest individuals have a greater probability of generating a new offspring, who, in turn, will generate another a new one and so on. To determine the fittest individual, we calculate the sum of distance among all points and its cluster centroid in each one of the three clusters. If this distance is lower in an individual than in others, it means the data inside that cluster are well organized, i.e., there are more points closer to its central point in a cluster than in others. For our purpose, the exchange of chromosomes will improve the solution, where we are finding the shortest total distance in a chromosome.

Each chromosome also undergoes a mutation probability, which is a fixed number. Mutation allows the beginning and preservation of genetic variation in a population by introducing another genetic structure modifying some gene inside the chromosome. The new mutated chromosome will be used to generate a new offspring.

The best population will be acquired at the end of these processes, and from this we choose the best individual, which will then represent the shortest sum of distance between each point in the cluster and its respective centroid. So, we calculate the average among the three cluster centroids. This number represents a single point in the graphic, and this process will repeat for another 288 times, which represent all 5 minutes intervals during a day. By using data from three previous days to generate this single point, we now have a network signature of this day, or the DSNSF-GA.

IV. Tests and Results

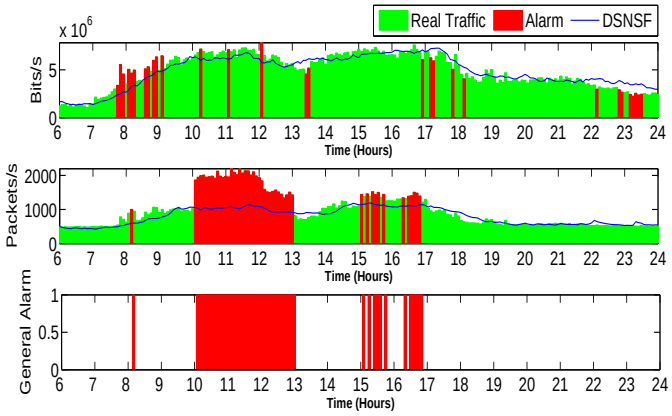
As described before, we used real information obtained from the historical database of the State University of Londrina (UEL). We generated the DSNSFs for the period of two weeks. Furthermore, we can see from Figure 3 the alarms generated by the change on traffic behavior. These alarms are clear during DDoS and DoS attacks artificially generated using the Scorpius software [25]. Basically, this tool injects abnormal flows directly into the exported real data flows according to the specific behavior of the desired anomaly. We have set an interval between 10:00 and 13:00 for the DDoS attack and between 15:00 and 17:00 for DoS attack for the 23^{rd} April. As the UEL working hours are from 07:00 to 23:00 hours, the historical database were analyzed for the period between 06:00 and 24:00 hours. The DSNSFs are presented in Figures 1 and 2 where the green color represents the real traffic, the red line represents the DSNSF-FA and the blue line the DSNSF-GA, both indicating the expected traffic according to their rules. The first week analyzed were from 22^{nd} to 26^{th} April 2013 and the second from 29^{th} April to 3rd May 2013.

The key process for an anomaly detection system is the traffic characterization. Both methods work characterizing traffic from sFlow data, each one using a different metaheuristic technique. Based on that traffic depiction, we can compare the prediction and the real traffic and identify the anomaly. Our intent is to compare both methods. To evaluate the accuracy of our models for these two weeks, three metrics were used: the Correlation Coefficient, the Normalized Mean Squared Error (NMSE) and the ROC curve [26].

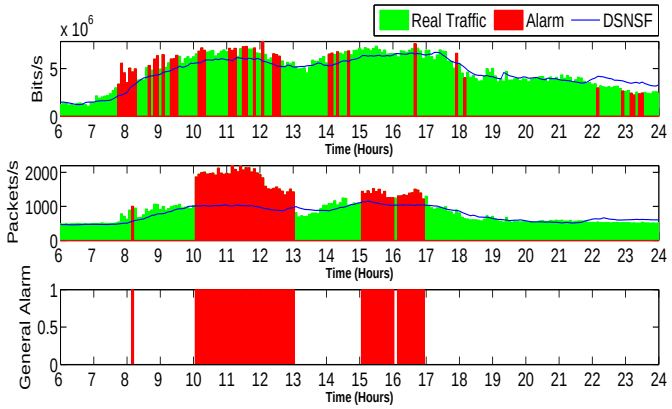
The Correlation Coefficient (CC) function is to indicate the direction and strength of the relationship between two variables (for our propose, each DSNSF and the real data of the day). In other words, if the changes suffered by a variable are accompanied by the other, there is a correlation between them. CC has its value $\in [-1, 1]$, where 1 indicates strong positive correlation, -1 strong negative correlation and 0 corresponds to no correlation. Each week are shown in the Tables I and II.

In Tables I and II, according to the averages, both models showed good results with strong correlation in normal days, where CCs are very close to 1, and the differences found between the DSNSF-FA and DSNSF-GA were small. For the 23^{rd} April, we can see small values, both for FA and GA, specially when packets per second were analyzed. When bits per second were analyzed, there was no difference for CC.

 ISBN: 978-1-61208-411-4



(a) DSNSF-FA.



(b) DSNSF-GA.

Figure 3. DSNSFs Alarms for 23^{rd} April

Also, we found two other abnormal values. One from the 1st May 2013 caused by a national holiday, where we had few activities in the UEL and another for 25^{th} April. We have here a classical flash crowd traffic, caused by students applying for their enrollment in the Business Administration course, being this the last day for enrollment and only available via the Internet, where the web serves are located inside the UEL network.

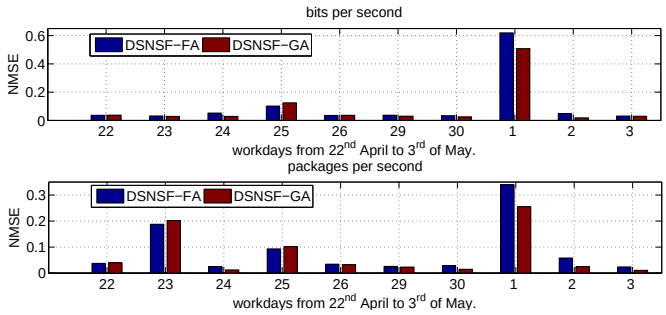


Figure 4. NMSE to DSNSF-FA and DSNSF-GA.

The NMSE is the mean square of the difference between analyzed values, checking the model's predictive ability. Their values are for $0 \leq NMSE \leq 1$, and values closer to zero are the most faithful DSNSF. Figure 4 illustrates the NMSEs results for bits and packets per second, obtained by the models. Note that both DSNSF-FA and DSNSF-GA managed NMSE values below 0.02 in most days. For 23^{rd} April we found a high value for packets per second again, obviously caused by the injected attacks, which confirms that our models are able to identify deviations. Also, due to abnormal traffic on 1^{st} May 2013 caused by the national holiday, and for 25^{th} caused by the students enrollment, we found high values, both for packets and bits per second.

TABLE I. CC TABLES - DAYS BETWEEN 22^{nd} to 26^{th} OF APRIL 2013

CC\Days	22	23	24	25	26	Average
FA-bits	0.88	0.88	0.85	0.78	0.87	0.85
FA-Packets	0.87	0.74	0.86	0.64	0.82	0.81
GA-bits	0.91	0.91	0.92	0.77	0.89	0.88
GA-Packets	0.93	0.80	0.92	0.69	0.87	0.86

TABLE II. CC TABLES - DAYS BETWEEN 29^{th} OF APRIL TO 1^{st} OF MAY 2013

CC\Days	29	30	1	2	3	Average
FA-bits	0.88	0.87	0.36	0.85	0.88	0.77
FA-Packets	0.87	0.85	0.15	0.81	0.85	0.79
GA-bits	0.94	0.92	0.49	0.93	0.91	0.84
GA-Packets	0.93	0.91	0.16	0.88	0.88	0.84

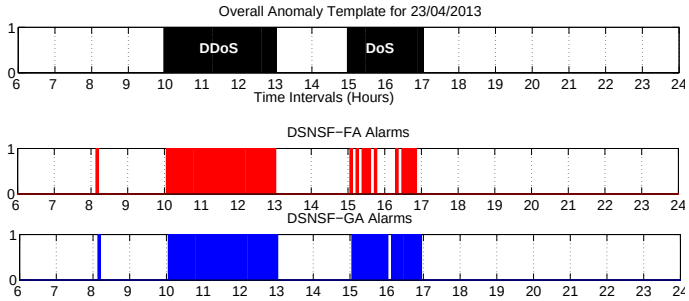


Figure 5. General alarm comparison.

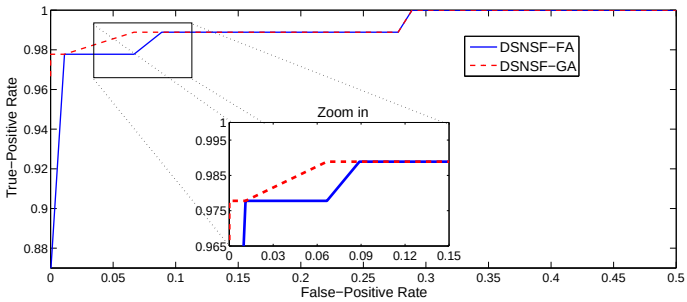


Figure 6. ROC curves comparing the trade-off between TPR and FPR rates of the proposed methods.

Figure 5 shows DDoS e DoS attacks artificially inserted and the alarms generated by the models. The data that triggered these true and false alarms, obtained by the technique of Adaptive Dynamic Time Warping (ADTW)[27], called true-positive rate (TPR) and false-positive rate (FPR) respectively, were used as the basis for the curve construction of the Receiver Operating Characteristic (ROC) and the extent accuracy of both models.

In DDoS attack's detection, both models obtained 97.3%. Moreover, for DoS attack the DSNSF-FA obtained 48% and the DSNSF-GA 88%.

 ISBN: 978-1-61208-411-4

The ROC curve, presented in Figure 6, describes the trade-off between TPR and FPR, which allowed to obtain the performance of DSNSF-FA and DSNSF-GA on the detection of generated artificial abnormalities. Analyzing the figure's zoom in, we notice that both models had a great performance with a minimum detection of false alarms. DSNSF-GA had a trade-off of 93.5% TPR with 0.4% FPR, as DSNSF-FA reaches 77.4% TPR with 0.4% FPR. Concerning the accuracy measure, DSNSF-GA had an accuracy of 98.3% and DSNSF-FA obtained 94.8%. The efficiency measure of the models were 96.5% to DSNSF-GA and 88.5% to DSNSF-FA.

V. Conclusion

In this work, we used two metaheuristics to create a Digital Signature of Network Segment using Flow Analysis (DSNSF). The first model uses FA to generate the DSNSF using data such as bits and packets per second, collected using sFlow pattern from the State University of Londrina (UEL). The second model uses GA to generate the DSNSF using the same set of data. Both models work characterizing traffic and comparing the predicted with the real traffic. In addition, we injected anomalous traffic in a specific day to analyze its behavior and evaluate the results to measure the efficiency of our models, finding good results.

We could see in the tables and graphs provided that both models are able to identify anomalous traffic using data such as bits and packets per second with a small advantage for the DSNSF-GA model, specially when we consider the number of true-positive alarms for DoS attacks, due to the efficiency measure and the accuracy. For future works, we intend to increase the number of dimensions in our search, since network flows can give us more data, such as IP and ports information for example.

Acknowledgment

This work was supported by SETI/Fundação Araucária and MCT/CNPq for Betelgeuse Project financial support. Also, the authors would like to thank the support from São Paulo State Technological College (Fatec Ourinhos).

References

[1] E. Petac, A. Alzoubaidi, and P. Duma, "Some experimental results about security solutions against ddos attacks," in Signals, Circuits and Systems (ISSCS), 2013 International Symposium on, July 2013, pp. 1–4.

[2] B. Trammell and E. Boschi, "An introduction to ip flow information export (ipfix)," IEEE Communications Magazine, vol. 49, no. 4, 2011, pp. 89–95.

[3] P. Teodoro, P. Feldstedt, and D. Zuiga, "Automatic signature generation for network services through selective extraction of anomalous contents," in Telecommunications (AICT), 2010 Sixth Advanced International Conference on, May 2010, pp. 370–375.

[4] A. H. Gandomi, X.-S. Yang, and A. H. Alavi, "Mixed variable structural optimization using Firefly Algorithm," Computers & Structures, vol. 89, no. 23-24, Dec. 2011, pp. 2325–2336.

[5] T. Hassanzadeh, H. Vojodi, and A. M. E. Moghadam, "An image segmentation approach based on maximum variance Intra-cluster method and Firefly algorithm," in 2011 Seventh International Conference on Natural Computation. IEEE, Jul., pp. 1817–1821.

[6] J. H. Holland, Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. U Michigan Press, 1975.

[7] U. Maulik and S. Bandyopadhyay, "Genetic algorithm-based clustering technique," Pattern Recognition, vol. 33, no. 9, 2000, pp. 1455–1465.

[8] M. L. Proença Jr., B. B. Zarpelão, and L. S. Mendes, "Anomaly detection for network servers using digital signature of network segment," in Proceedings - Advanced Industrial Conference on Telecommunications/Service Assurance with Partial and Intermittent Resources Conference/E-Learning on Telecommunications Workshop AICT/SAPIR/ELETE 2005, vol. 2005, 2005, pp. 290–295, doi:10.1109/AICT.2005.26.

[9] R. Steinert and D. Gillblad, "Towards distributed and adaptive detection and localisation of network faults," in Telecommunications (AICT), 2010 Sixth Advanced International Conference on, May 2010, pp. 384–389.

[10] M. V. O. Assis, J. J. P. C. Rodrigues, and M. L. Proença Jr., "A seven-dimensional flow analysis to help autonomous network management," Information Sciences, vol. 278, 2014, pp. 900 – 913, doi:10.1016/j.ins.2014.03.102.

[11] E. H. M. Pena, S. Barbon, J. J. P. C. Rodrigues, and M. L. Proença Jr., "Anomaly detection using digital signature of network segment with adaptive arima model and paraconsistent logic," in Computers and Communication (ISCC), 2014 IEEE Symposium on, June 2014, pp. 1–6, doi:10.1109/ISCC.2014.6 912 503.

[12] M. L. Proença Jr., C. Coppelmans, M. Bottoli, and L. Souza Mendes, "Baseline to help with network management," in e-Business and Telecommunication Networks. Springer Netherlands, 2006, pp. 158–166, doi: 10.1007/1–4020–4761–4_12.

[13] B. B. Zarpelão, L. S. Mendes, M. L. Proença Jr., and J. J. P. C. Rodrigues, "Parameterized anomaly detection system with automatic configuration," in Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE, Nov 2009, pp. 1–6, doi: 10.1109/GLOCOM.2009.5 426 189.

[14] A. A. Amaral, B. B. Zarpelão, L. de Souza Mendes, J. J. P. C. Rodrigues, and M. L. Proença Jr., "Inference of network anomaly propagation using spatio-temporal correlation," Journal of Network and Computer Applications, vol. 35, no. 6, 2012, pp. 1781 – 1792, doi: 10.1016/j.jnca.2012.07.003.

[15] P. Phaal, S. Panchen, and N. McKee, "InMon corporation s sFlow: A method for monitoring traffic in switched and routed networks," RFC 3176, Tech. Rep., 2001.

[16] Z. Güngör and A. Ünler, "K-harmonic means data clustering with simulated annealing heuristic," Applied Mathematics and Computation, vol. 184, no. 2, Jan. 2007, pp. 199–209.

[17] X.-S. Yang, "Firefly algorithms for multimodal optimization" , in: Stochastic Algorithms: Foundations and Applications, ser. Lecture Notes in Computer Science, O. Watanabe and T. Zeugmann, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, vol. 5792.

[18] I. Fister, X.-S. Yang, and J. Brest, "A comprehensive review of firefly algorithms," Swarm and Evolutionary Computation, Dec. 2013, pp. 34–46.

[19] J. Macqueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1 (Univ. of Calif. Press, 1967), 281-297, vol. 1, 1967, pp. 281–297.

[20] I. Paschalidis and G. Smaragdakis, "Spatio-temporal network anomaly detection by assessing deviations of empirical measures," Networking, IEEE/ACM Transactions on, vol. 17, no. 3, June 2009, pp. 685–697.

[21] M. H. A. C. Adaniya, M. F. Lima, J. J. P. C. Rodrigues, T. Abrao, and M. L. Proença Jr., "Anomaly detection using dsns and firefly harmonic clustering algorithm," in Communications (ICC), 2012 IEEE International Conference on, June 2012, pp. 1183–1187, doi:10.1109/ICC.2012.6 364 088.

[22] P. R. G. Hernandes Jr, L. F. Carvalho, G. Fernandes Jr., and M. L. Proença Jr., "Digital signature of network segment using genetic algorithm and ant colony optimization metaheuristics," in The Eighth International Conference on Emerging Security Information, Systems and Technologies, Nov 2014, pp. 62–67.

[23] "nfdump - documentation," http://nfdump.sourceforge.net/, 2014, access date: May 13, 2015.

[24] K. Wang, B. Wang, and L. Peng, "CVAP: Validation for Cluster Analyses," Data Science Journal, vol. 8, 2009, pp. 88–93.

[25] "Scorpius - sflow anomaly simulator," http://redes.dc.uel.br/scorpius/, 2013, access date: Apr 28, 2015.

[26] B. Penney, M. King, and S. Glick, "Restoration of combined conjugate images in spect: comparison of a new wiener filter and the image-dependent metz filter," Nuclear Science, IEEE Transactions on, vol. 37, no. 2, Apr 1990, pp. 707–712.

[27] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 26, no. 1, Feb 1978, pp. 43–49.

 ISBN: 978-1-61208-411-4

Discovering Attack Strategies Using Process Mining

Sean Carlisto de Alvarenga, Bruno Bogaz Zarpelão, Sylvio Barbon Junior
Computer Science Department
State University of Londrina (UEL)
Londrina, Paraná, Brazil
E-mail: {sean, brunozarpelao, barbon}@uel.br

Rodrigo Sanches Miani
School of Computer Science (FACOM)
Federal University of Uberlândia (UFU)
Uberlândia, Minas Gerais, Brazil
E-mail: miani@ufu.br

Michel Cukier
A. James Clark School of Engineering
University of Maryland
College Park, Maryland, USA
E-mail: mcukier@umd.edu

***Abstract*— Intrusion Detection Systems generate alerts which depend on manual analysis of a specialist to determine a response plan. However, these systems usually trigger thousands of alerts per day. Investigating unmanageable amounts of alerts manually becomes burdensome and error-prone. Besides, it complicates the analysis of critical alerts. In this paper, an approach is proposed to facilitate the investigation of huge amounts of intrusion detection alerts by a specialist. The proposed approach makes use of process mining techniques to discover attack strategies observed in intrusion alerts, which are presented to the network administrator in friendly visual models. Tests were performed using a real dataset from the University of Maryland. The results show that the proposed approach combines visual features along with quantitative measures that help the network administrator to analyze the alerts in an easy and intuitive manner.**

Keywords-intrusion detection; security visualization; alert mining; heuristic mining.

I. INTRODUCTION

In recent years, the increase of security vulnerabilities has concerned companies and organizations. In 2014 alone, almost 8000 new vulnerabilities were found in software applications and operating systems, as shown by the National Vulnerability Database (NVD) statistics [1]. The more increases the number of new vulnerabilities, the greater the likelihood of increase in the frequency of computer security violations. That is where the security measures come into play.

Intrusion Detection Systems (IDS) are devices that play an important role in the set of security policies in information systems. IDS monitor the network and system activities for any security violations. When it detects a security violation, it reports the event to a network administrator, who assesses the threat and initiates a response [2]. Unfortunately, IDS sensors generate huge amounts of alerts that makes it difficult to analyze them and identify relevant alerts [3]. To address this problem, alert correlation techniques [3][4][5] have been proposed to extract high-level descriptions of huge amounts of alerts.

The idea of using high level descriptions and graphical models in security assessment is not exclusive of alert correlation research, but it is also employed in the theory of attack trees and attack graphs. Attack trees and attack graphs have been extensively used to a variety of purposes such as attack and defense assessment, as well as for metrics quantification (e.g., cost, time, impact, probabilities, etc.). However, these representations usually require some expert knowledge of the network (e.g., topology, hosts) to generate the model.

In this paper, an approach is proposed to the IDS alert analysis problem from a process-oriented perspective. Alerts are considered as events of a process and they are analyzed with process mining techniques to generate a process model. The process model is a high-level visual representation of attack strategies observed in IDS alerts.

The proposed approach has the following benefits. At first, specific data acquisition is not necessary since companies and organizations usually employ IDS sensors to protect their networks. Secondly, process models provide an intelligible and intuitive way to interpret complex information such as IDS alerts. Thirdly, it is possible to model different perspectives from the alerts, e.g., the attackers' perspective, giving the network administrator a comprehensive view of the network. Moreover, it supports different levels of granularity in analysis as it is possible to filter the most frequent behavior observed in the alerts. Finally, the proposed approach shows the strategies that attackers are employing to compromise the network, helping network administrators to determine preventive measures.

The rest of the paper is organized as follows. Section II reviews related work. Section III defines the preliminary concepts used in this paper. Section IV shows the proposed approach and its operation. Section V presents the results obtained in the evaluation of the proposed approach. Finally, the Section VI contains concluding remarks and future work possibilities.

 ISBN: 978-1-61208-411-4

II. RELATED WORK

In this section, an overview of previous work on attack modeling and IDS alert analysis is presented. Previous work that used real IDS alerts data to discover attack strategies with process mining techniques was not found in literature. Therefore, approaches that use visual representation for attack modeling and data mining for IDS alerts analysis will be presented.

One of the great advantages of using higher level graphical models is that they are intuitive and facilitate threat assessment and attack scenario understanding. Attack trees and attack graphs are the most common methods used to modeling attack threats. As introduced by Schneier [6], attack trees are a visual representation that aims at modeling an attack in a tree structure. The attacker's goal is specified as the root of the tree. Branches in the tree represent attack subgoals, which can be represented as disjunctive or conjunctive nodes. Disjunctive nodes depict different alternative paths that an attacker can follow to achieve his goal. Conjunctive nodes represent different steps an attacker needs to take in order to achieve a goal [7].

Unlike the approach proposed in this work, attack trees are often modeled manually, a labor-intensive and error-prone process. In [8][9][10], this problem is addressed by methods to automate attack trees generation. Moore et al. [11] use attack trees to represent security attacks and document information, aiding security analysts to identify attack patterns. Tidwell et al. [12] enhance attack trees to represent multi-stage attacks behavior with an attack specification language.

Attack trees have some limitations regarding attacks modeling. This type of representation is static and can not take temporal aspects, such as dynamic time variations and order or priority of actions [7]. Therefore, this representation is not suitable for the proposed approach that takes these aspects into account.

Attack graphs are another way to represent and analyze security attacks. The term was first introduced by Phillips and Swiler [13]. In an attack graph, the nodes represent the network state and the edges represent an action of the attacker that changes the state. Weights can be assigned to the edges to enrich the model and algorithms can be applied to graph analysis, e.g., shortest path, to find which paths are more likely to succeed, time to success and other metrics [7]. Swiler et al. [14] developed a tool to generate attack graphs. Researches in [15][16] addressed the scalability problem of the graph size. Attack graphs are generated based on information about the attack, the system and the attacker profile [7]. This requires some background knowledge that is not always known. The approach proposed in this work generates the model based only on IDS alerts and hence does not require such knowledge.

Researchers have also studied how to extract attack information from huge volumes of IDS alerts. In [3], Ning and Xu published one of the first researches in this field. They proposed a model that builds graphs from IDS alerts to represent attack strategies. The authors also presented a method to measure the similarity between different attack strategy graphs. In more recent work, Lagzian et al. [4] and Xuewei et al. [5] used data mining techniques. Lagzian et al. presented a framework that, at first, aggregates the alerts in graphs. Then, it applies the Bit-AssocRule algorithm to mine the most frequent patterns in the graphs. Xuewei et al., on the other hand, proposed to identify causal relationships between the alerts with Markov models.

III. BACKGROUND INFORMATION

A. Intrusion Detection Systems

An IDS is a software or a hardware device that monitors computers or network traffic for malicious activities or intrusive behavior. Once a malicious activity is detected, IDS can either raise an alert or log the event [17]. IDS can be classified into two categories, namely network-based and host-based. Moreover, it can use one of these three techniques: signature-based detection, anomaly-based detection or hybrid [18].

Signature-based detection is the process of comparing patterns or signatures that corresponds to a known threat against observed network events to identify malicious activity. This technique uses a database of already known attack signatures for detecting intrusions. Signature-based IDS is very effective to detect known attacks, already defined in its database. On the other hand, it can not detect attacks that do not have a previous signature, e.g., zero-day attacks or modified attacks. This limitation is circumvented by adding new signatures and keeping the database up to date.

An anomaly-based IDS works by distinguishing an abnormal behavior from what is considered to be normal. Therefore, this technique builds a model of normal traffic and raises an alert for any traffic that deviates from this model. A great advantage of this method is the detection of new attacks without any prior knowledge. The weakness of anomaly detection is the difficulty to define a model for what is normal, what is malicious and the boundaries between them.

A hybrid method combines the qualities of both signature-based and anomaly-based detection and integrates them in a single system.

B. Process Mining

Process mining depicts a set of methods and approaches that combine data mining techniques and business process modeling and analysis [19]. Process mining uses information recorded in a log to extract knowledge and represent it as process models. Therefore, it is important that logs have relevant and proper information as they are the starting point for process mining techniques.

For process mining, each record in the log is considered an event, the reason the logs are known as event logs. Furthermore, to extract information from the event logs, some characteristics must be considered [20]:

- Each event in the log corresponds to an *activity*, i.e., an action that was performed in the process [20]. As an example, suppose a user registration system that records all its actions in a log. Each recorded action,

 ISBN: 978-1-61208-411-4

e.g., *Create User*, *Update User*, *Delete User*, etc., can represent an activity in the process.

- Each event in the process has to refer to a process instance or *case*. A *case* defines the process scope, i.e., where the process starts and where it ends. In the example of a user registration system, a set of events associated to the registration of a particular user can compose a *case*.
- Events can have attributes such as *activity*, *time* and *resource*. The attribute *activity* shows the event action, as mentioned before. The attribute *time* records the event timestamp. Finally, the attribute *resource* presents the responsible for performing the event.
- Events within a case are ordered as they occur, e.g., according to their timestamp. The occurring sequence of events is crucial because process mining algorithms determine causal dependencies between events to build the model.

There are three main areas in process mining, namely process discovery, conformance checking and model enhancement. Process discovery is related to how to transform an event log into a model. A process discovery technique receives as input an event log and returns as output a process model, so the model is representative for the behavior observed in the event log [19]. This is the main focus of the approach proposed in this work. Conformance checking uses metrics such as fitness and precision to evaluate the process model in the context of a log. Model enhancement uses new information to improve the process model.

In the next subsections, some process discovery techniques are briefly discussed. It is out of scope to discuss in details how these algorithms work, but benefits and drawbacks of each one will be pointed out. Further details can be found in [19]-[26].

C. The α-Algorithm

Proposed by van der Aalst et al. [23] in 2003, the α-Algorithm is one of the first algorithms designed for process mining and its ideas contributed to the development of more powerful discovery algorithms currently in use. The algorithm produces as output a WorkFlow net (WF-net), which is a subclass of Petri nets. In a WF-net, all nodes are on a path from the source place (unique place where the process starts) to the sink place (unique place where the process ends). The α-Algorithm examines the event log for four ordering relation between activities: directly follows relation, dependency relation, non-parallel relation and parallel relation. Refer to [22][23] for a complete description of the algorithm.

Under some specific conditions, the α-Algorithm works well. However, it has problems to deal with some situations (control-flow constructs) found in real life event logs. For instance, short loops, i.e., loops of length one or two, make the algorithm to derive an incorrect WF-net. Short loops occur when the same activity or two activities are executed multiple times in sequence. Considering IDS alerts, this may happen when the attacker attempts to perform the same violation several times until succeed.

The α-Algorithm has other limitations as it may not derive a correct WF-net when dealing with noise, i.e., event log with rare events that do not represent the process behavior, and incompleteness, i.e., the event log does not have enough events to discover a model. Therefore, this technique is not suitable in most real life processes.

D. The α-Algorithm extensions

To overcome the α-Algorithm limitations, many extensions have been proposed. Each of them extends the α-Algorithm to add support to some constraint. The α^{+}-Algorithm deals with the short loop problem. The Tsinghua-α-Algorithm focuses on event logs containing activities associated to transactional life-cycle. The α^{++}-Algorithm seeks to support non-free-choice control-flow construct. The $\alpha^{\#}$-Algorithm and the α^{*}-Algorithm concentrates on discovering some Petri nets that are not in the class of WF-nets and hence can not be discovered by the basic algorithm. Refer to a survey in [26] for more details.

E. Heuristic Mining

As mentioned in Section III-C, one of the limitations of the α-Algorithm is it can not deal with noise. However, noise is common in real life event logs due to information incorrectly logged and occurrence of exceptional events [23]. The Heuristic Mining algorithm handles this problem by taking the frequencies of events into account. Therefore, the algorithm can express the main behavior observed in the log without including the low frequency behavior from the noise into the model. Short loops are also overcome by the use of dependency/frequency table (D/F-table) and the dependency score [25]. The D/F-table contains metrics about the frequency of ordering relations occurrence, e.g., number of times one activity is directly followed by another activity. Based on these metrics, the dependency score, a numeric value between -1 and 1, is computed. The dependency score represents how strong the dependency relation between activities is. For instance, if the dependency score between activity *a* and itself is close to 1, then *a* is often the cause of *a*, suggesting a loop. These metrics along with dependency score and a threshold can be used to refine the output model.

IV. PROPOSED APPROACH

In this section, the proposed method to automate the discovery of attack strategies using a process mining discovery algorithm will be introduced. The proposed approach consists of four steps. In the first step, alerts with common features are aggregated. In the second step, the aggregated alerts are converted in a suitable format for process discovery algorithms. In the third step, the process discovery algorithm is executed to build the attack model. Finally, in the last step, the resulting attack model is analyzed. Figure 1 shows the four steps that compose the proposed approach.

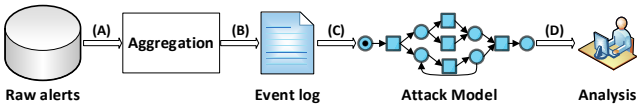


Figure 1. The four steps of the proposed approach.

In the following subsections, the details of each step are described. Then, in the next section, the method is evaluated

using a real dataset of IDS alerts, discussing the results and some considerations.

A. IDS alerts aggregation

In attack strategy discovery, the goal is to discover how the attackers are attempting to compromise the network. After attack strategy discovery, network administrators can know each step attackers often take and the dependencies between these steps. Therefore, in the first stage of the proposed approach, alerts with common features are aggregated, aiming to group the alerts that compose each attack strategy. Then, in the next stage, the process discovery algorithm will investigate the relationships between these alerts. To aggregate the alerts, the perspective to be represented is taken into account. The perspective denotes how the alerts will be associated in the aggregation process. For instance, to represent the attackers' perspective, one can aggregate the alerts originating from the same source IP address. Similarly, to represent the targets' perspective, one can aggregate the alerts with the same destination IP address. The flexibility to represent different perspectives in this step can be explored to provide the network administrator a comprehensive view of the network.

B. Conversion of aggregated alerts to an event log

As aforementioned, for process mining, the input dataset should consist of events recorded in a log. Therefore, since the intention is to use process mining in IDS alerts analysis, the second step of the proposed approach is to convert the aggregated alerts into an event log. IDS collect information that may vary according to the type of device. This information may include source IP address, destination IP address, source port, destination port, Autonomous System Number (ASN) information, signature severity, attack type group for each signature, etc.

To analyze IDS alerts under a process mining perspective, each individual alert is considered an event. Each event attribute (i.e., the attributes of the alerts) will be analyzed to build the attack model. Because the objective is to discover attack strategies, the alert attributes that provide information about the attacks must be chosen. Then, this information will be used to build an event log with the characteristics required by process mining such as the concepts of *case*, *activity* and *time* (see Section III-B).

At first, event activity (i.e., the action performed in the process) has to be defined. The event activity is an important information as it will be denoted by the nodes in the attack model. The nodes in the model represent the steps performed in the attack-flow and help the identification and visualization of sequences and dependencies of attacks in the model. Usually, IDS record information about what triggered the alert, e.g., some signature identification or description of the violation. In the context of IDS alerts, the signature can be considered the action of the attacker as it depicts his intentions to compromise the network. Therefore, the signature is defined as the event activity.

Moreover, in an event log, events should be grouped in a case (i.e., each event in the process belongs to a case). The case defines the scope of the process. During the process discovery, several cases are compared among each other to determine the causal dependencies between activities. In the proposed approach, a case is defined as a group of alerts that were aggregated in the first step (see Section IV-A) and occurred within a time span *t*. As an example, suppose that the alerts are aggregated according to the source IP address and the time span *t* is set as 1 day. Then, all alerts with source IP address *x.x.x.x* triggered in day *m* will belong to case *i*. All alerts with source IP address *x.x.x.x* triggered in day *n* will belong to case *j*. Finally, all alerts with source IP address *y.y.y.y* triggered in day *m* will belong to case *k*. In this manner, each attacker composes a case and his attack steps (i.e., its alerts occurred within *t*) are the events of the case. Finally, in an event log, events in a case must be ordered as they occur. In the IDS alerts context, the timestamp information is used to order the alerts.

The event log format adopted in the proposed approach is the eXtensible Event Stream (XES). XES is an eXtensible Markup Language (XML)-based standard used to store event logs supported by most process mining tools including the ProM Framework [27] used in this research.

C. Attack model discovery

To build the model, the process discovery algorithm that will take the event log as input and generate the attack model as output must be defined. As mentioned before, process discovery algorithms have limitations regarding the control-flow constructs they can discover. Different algorithms may generate different attack models. Furthermore, some algorithms may generate attack models that are not able to represent the behavior observed in the event log and consequently may lead to wrong conclusions about the attacks.

In IDS alerts, loops may take place in the model, since events that compose a case may have repeated activities in sequence (e.g., situations in which the attacker executed the same violation until succeed or attempted to compromise multiples hosts such as in a *botnet*). The discovery algorithm should be able to detect these repeated activities and represent them not as individual activities in sequence but as a loop in the model. On the other hand, duplicate tasks (e.g., situations in which two different violations have the same signature) will unlikely be a problem because IDS alerts are atomic entities (e.g., a *buffer overflow* exploit will not have the same signature of a *nimda attack* in the log). Therefore, the proposed approach uses the Heuristic Mining algorithm as it can deal with these characteristics.

D. Model evaluation

After the model has been generated, an expert analysis is required. Through the model, different aspects can be observed. In the next section, a case study and some analysis will be presented and discussed.

 ISBN: 978-1-61208-411-4

V. RESULTS

To evaluate the proposed approach, raw IDS alerts generated by a signature-based device deployed at the University of Maryland, whose network has about 40000 computers, were used. These alerts were triggered between April and December 2012 for inbound and outbound network traffic of the University. The alerts raised in October were chosen to evaluate the method.

To perform the first step of the proposed approach, alerts with the same source IP address were aggregated. Then, the cases were defined, setting the time span *t* to 1 day. Consequently, alerts with the same source IP address that were triggered in the same day were associated to the same case. Moreover, only inbound alerts, i.e., alerts originating from traffic addressed to the University were considered. To represent the frequent behavior of the attackers, exceptional situations were filtered (similar to the filtering performed in [3]). Therefore, cases containing a single event (isolated alert) or cases containing multiple events associated to the same violation (i.e., same signature) were not included in the model. These cases do not depict attack strategies used by attackers, as they show an attack-flow with a single step and do not provide enough information on the behavior of the attacker, as illustrated by Figure 2.

Similarly, cases containing more than 50 events in which almost all the events have the same signature were not included in the model.

The ProM Framework [27] was used to generate the process model with the Heuristic Mining algorithm. Figure 3 shows the results of tests performed on October 7th. In this day, there were 97 events (i.e., triggered alerts) with 8 different activities (i.e., distinct signatures) organized in 9 cases.

Analysis of Figure 3 indicates that:

- Within that day, the attacks started in one of four violations: (i) *Malicious PHP Program Access*, (ii) *Malicious SMB Probe/Attack*, (iii) *Possible nmap Scan (XMAS (FIN PSH URG))* and (iv) *Impossible Flags (SFRPAU)*. Each of them lead to a different attack-flow.
- Among the 9 cases, there is one case that starts with *Malicious PHP Program Access*, two cases that start with *Malicious SMB Probe/Attack*, three cases that start with *Possible nmap Scan (XMAS (FIN PSH URG))* and three cases that start with *Impossible Flags (SFRPAU)*. Similarly, one case ends with *PHP Code Injection*, one case ends with *Windows PlugnPlay Request Anomaly*, four cases end with *Possible nmap Scan (XMAS (FIN PSH URG))* and three cases end with *Impossible Flags (SFRPAU)*.
- In (i), an attack can be clearly observed. First, the attacker executes a *Malicious PHP Program Access*. Afterwards, the attacker executes a *PHP Code Injection* and then the two activities come into loop. This shows that some attacker is injecting code (e.g., eval injection) into a PHP server located at the University network and then some user/visitor is accessing the server and executing the code. This attack-flow shows a possible unknown vulnerability that the network administrator has to fix.
- In (ii), in one of the attack-flows, the attacker performs a *Malicious SMB Probe/Attack* followed by *Windows PlugnPlay Request Anomaly*. Although not directly related, both attacks have something in common: they are associated to Microsoft Operating System (OS) and exploit vulnerabilities that allow remote code injection and elevation of privileges. These vulnerabilities, if successfully exploited, can allow the attacker to take control of the compromised system as reported by Microsoft Security Bulletin [28][29].
- In (iii), a possible attack attempting is presented. The attacker performs a port scan (*TCP Xmas scan*), probing the server or host for open ports. Port scan is a well known technique used in pre-attack phases to gather information about the target and be able to exploit them. After the port scan, the attack-flow splits into three paths. One path leads to (ii). The other path leads to (iv). In the third path, the attacker performs *NULL OS Fingerprinting Probe*, an attempt to collect information about the target OS and thus know what vulnerabilities he can/can not exploit (e.g., if the vulnerability was already patched in this OS version). After that, the path leads to (iv). This attack-flow indicates that the attacker is conducting a reconnaissance of the target before executing the attack.
- In (iv), the attack-flow is similar to (iii). The *Impossible Flags (SFRPAU)* are TCP packets with all flags (*SYN*, *FIN*, *RST*, *Push*, *ACK*, *UrgPtr*) set. These packets might be unintentional produced by poorly implemented applications but are more likely (considering the attacks in the paths it splits) from a *Full Xmas scan*.

It is possible to obtain other information by analyzing the model. For example, the *Impossible Flags (SFRPAU)* signature was the most executed attack (30 times). Next, there is the *Possible nmap Scan (XMAS (FIN PSH URG))* attack (21 times). The reason for this behavior is the loop between the attacks, showing that many port scans were executed in this day. In addition, the model provides an intuitive and easy way to investigate the alerts, showing the attack strategies that would hardly be discovered investigating almost 100 alerts manually.

As mentioned before, the attack model presented in Figure 3 represents the attackers' perspective, i.e., how multiple source IP addresses (i.e., the attackers) are attempting to compromise several targets in the University network. However, this representation may not be the ideal for all situations and other perspectives can be explored for a deep

 ISBN: 978-1-61208-411-4

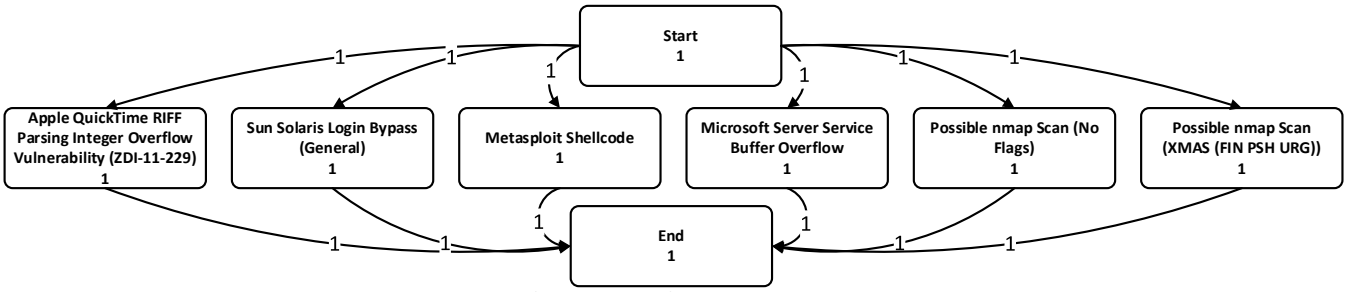


Figure 2. Example of the behavior of isolated alerts on October 7th.

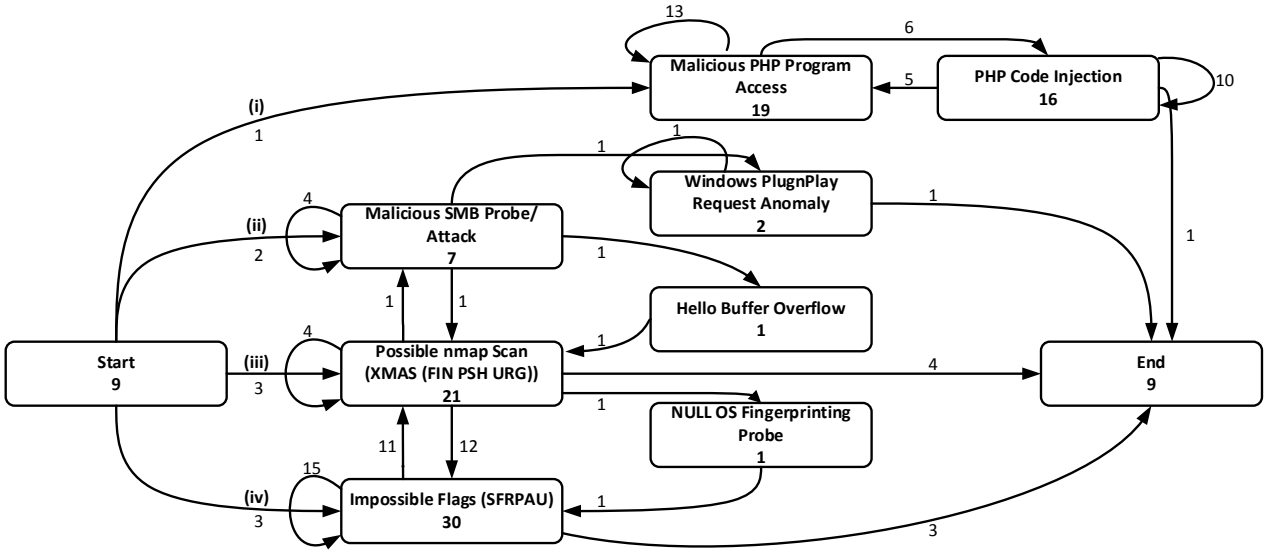


Figure 3. Attack model that represents the behavior of the attackers on October 7th.

investigation into the attacks. For instance, to represent the behavior of distributed attacks (many-to-one attacks), the targets' perspective can be explored (i.e., cases with events associated to the same destination IP address).

VI. Conclusion and Future Work

This paper has addressed the problem of analyzing huge amounts of IDS alerts. A four step method that uses process mining techniques to mine the alerts and generate a process model, a high-level graphical representation of the attackers' behavior observed in the alerts, was proposed. The method was evaluated on a real IDS alert dataset from University Maryland. The results showed that the resulting model has an intuitive and user-friendly representation that can be used by network administrators as an alternative to the manual investigation of alerts.

As future work, the objective is to extend the attack perspectives and analyze the alerts from another viewpoint (e.g., the target perspective). Besides, it was observed that some models become complex as the number of distinct signatures increases. Therefore, clustering techniques may be employed to reduce the complexity of those models and conformance checking metrics, such as simplicity, may be employed to evaluate the model.

Acknowledgment

The authors would like to thank Gerry Sneeringer and the Division of Information Technology at the University of Maryland for allowing and supporting the described research.

References

[1] "National vulnerability database," [Online]. Available: https://web.nvd.nist.gov/view/vuln/statistics. [Retrieved: April, 2015].

[2] S. O. Al-Mamory and H. Zhang, "Intrusion detection alarms reduction using root cause analysis and clustering," Computer Communications, 2009, vol. 32, no. 2, pp. 419-430.

[3] P. Ning and D. Xu, "Learning attack strategies from intrusion alerts," in Proceedings of the 10th ACM Conference on Computer and Communications Security. ACM, 2003, pp. 200-209.

[4] S. Lagzian, F. Amiri, A. Enayati and H. Gharaee, "Frequent item set mining-based alert correlation for extracting multi-stage attack scenarios," in Telecommunications (IST), 2012 Sixth International Symposium on. IEEE, 2012, pp. 1010-1014.

[5] F. Xuewei, W. Dongxia, H. Minhuan and S. Xiaoxia, "An approach of discovering causal knowledge for alert correlating based on data mining," in Dependable, Autonomic and Secure Computing (DASC), 2014 IEEE 12th International Conference on. IEEE, 2014, pp. 57-62.

[6] B. Schneier, "Attack trees: Modeling security threats," Dr. Dobb's Journal, December 1999. [Online]. Available:

 ISBN: 978-1-61208-411-4

https://www.schneier.com/paper-attacktrees-ddj-ft.html. [Retrieved: April, 2015].

[7] B. Kordy, L. Piètre-Cambacédès and P. Schweitzer, "DAG-based attack and defense modeling: Don't miss the forest for the attack trees," Computer Science Review, 2014, vol. 13–14, pp. 1-38.

[8] R. Vigo, F. Nielson and H. R. Nielson, "Automated generation of attack trees," in Computer Security Foundations Symposium (CSF), 2014 IEEE 27th. IEEE, 2014, pp. 337-350.

[9] S. Paul, "Towards automating the construction & maintenance of attack trees: a feasibility study," in Proceedings First International Workshop on Graphical Models for Security, GraMSec 2014, 2014, pp. 31-46.

[10] H. Birkholz, S. Edelkamp, F. Junge and K. Sohr, "Efficient automated generation of attack trees from vulnerability databases," in Working Notes for the 2010 AAAI Workshop on Intelligent Security (SecArt), 2010, pp. 47-55.

[11] A. P. Moore, R. J. Ellison and R. C. Linger, "Attack modeling for information security and survivability," Technical Note CMU/SEI-2001-TN-001, Carnegie Mellon University, 2001.

[12] T. Tidwell, R. Larson, K. Fitch and J. Hale, "Modeling internet attacks," in Proceedings of the 2001 IEEE Workshop on Information Assurance and security, 2001, pp. 54-59.

[13] C. Phillips and L. P. Swiler, "A graph-based system for network-vulnerability analysis," in Proceedings of the 1998 Workshop on New Security Paradigms, 1998, pp. 71-79.

[14] L. P. Swiler, C. Phillips, D. Ellis and S. Chakerian, "Computer-attack graph generation tool," in DARPA Information Survivability Conference & Exposition II, 2001. DISCEX '01. Proceedings. IEEE, 2001, pp. 307-321.

[15] S. Jajodia, S. Noel and B. O'Berry, "Topological analysis of network attack vulnerability," in Managing Cyber Threats, 2005, vol. 5, pp. 247-266.

[16] X. Ou, W. F. Boyer and M. A. McQueen, "A scalable approach to attack graph generation," in Proceedings of the 13th ACM Conference on Computer and Communications Security, CCS '06, 2006, pp. 336-345.

[17] A. Patel, Q. Qassim and C. Wills, "A survey of intrusion detection and prevention systems," Information Management & Computer Security, 2010, vol. 18, pp. 277-290.

[18] J. Vacca, Computer and Information Security Handbook, Second Edition, Morgan Kaufmann, 2013.

[19] W. M. van der Aalst, Process Mining: Discovery, Conformance and Enhancement of Business Processes, Springer Science & Business Media, 2011.

[20] W. van der Aalst and C. Giinther, "Finding structure in unstructured processes: The case for process mining," in Application of Concurrency to System Design, 2007. ACSD 2007. Seventh International Conference on, 2007, pp. 3-12.

[21] R. P. J. C. Bose and W. M. van der Aalst, "Context Aware Trace Clustering: Towards Improving Process Mining Results," SDM, 2009, pp. 401-412.

[22] A. de Medeiros, W. van der Aalst and A. Weijters, "Workflow mining: Current status and future directions," in On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, 2003, pp. 389-406.

[23] W. van der Aalst, T. Weijters and L. Maruster, "Workflow mining: discovering process models from event logs," Knowledge and Data Engineering, IEEE Transactions on, 2004, vol. 16, no. 9, pp. 1128-1142.

[24] A. Weijters and J. Ribeiro, "Flexible heuristics miner (FHM)," in Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on, 2011, pp. 310-317.

[25] A. Weijters and W. van der Aalst, "Rediscovering workflow models from event-based data using little thumb," Integrated Computer-Aided Engineering, 2003, pp. 151-162.

[26] B. van Dongen, A. Alves de Medeiros and L. Wen, "Process mining: Overview and outlook of petri net discovery algorithms," in Transactions on Petri Nets and Other Models of Concurrency II, 2009, pp. 225-242.

[27] B. van Dongen, A. de Medeiros, H. Verbeek and A. Weijters, "The prom framework: A new era in process mining tool support," in Applications and Theory of Petri Nets 2005, 2005, pp. 444-454.

[28] Microsoft, "Microsoft Security Bulletin MS05-039: Vulnerability in plug and play could allow remote code execution and elevation of privilege (899588)," August 2005, [Online]. Available: https://technet.microsoft.com/library/security/ms05-039. [Retrieved: April, 2015].

[29] Microsoft, "Microsoft Security Bulletin MS11-019: Vulnerabilities in SMB client could allow remote code execution (2511455)," April 2011, [Online]. Available: https://technet.microsoft.com/library/security/ms11-019. [Retrieved: April, 2015].

 ISBN: 978-1-61208-411-4

Control Plane Design for a Content Streaming System with Dual Adaptation

Eugen Borcoci, Cristian Cernat, Radu Iorga
University POLITEHNICA of Bucharest
Bucharest, Romania
emails: eugen.borcoci@elcom.pub.ro, cristian.cernat@elcom.pub.ro, radu.iorga@elcom.pub.ro

***Abstract* — Content streaming delivery has been recently considered as an attractive solution for media distribution, based on light architectures, working on top of the current Internet Protocol (IP) technologies. Such a system is considered here, integrating functionalities such as content server initial selection based on multi-criteria algorithm and then media adaptation (using dynamic adaptive streaming) and/or server switching - during the media session. This work-in-progress contributes to identify the main design concepts for the Control Plane of the architecture and in particular for the monitoring of Quality of Services (QoS) and Quality and Experience (QoE), aiming to support both the server selection and in–session adaptation actions.**

Keywords — Content delivery, Dynamic Adaptive Streaming over HTTP, Monitoring, Server and Path selection.

I. INTRODUCTION

Recently, over-the-top (OTT) solutions are proposed and developed, for media/content delivery, where the services are delivered over the current Internet by an entity called Service Provider (SP) that is not directly responsible for the quality of the flows transmission to the end-user; users access is done via the "public Internet". The OTT SP could exist as a separate entity from traditional Internet Service Provider (ISP). Also, combined solutions exist, with OTT Service Providers using the Content Delivery Network (CDN) Providers' infrastructure to improve the quality of delivery.

A *light (OTT-like) novel architecture* for content streaming systems over the current Internet is proposed by the European *DISEDAN* Chist-Era project [3], (*service and user-based **DI**stributed **SE**lection of content streaming source and **D**ual **A**daptatio**N**,* 2014-2015). The business actors involved are: *Service Provider (SP)* - an entity/actor which delivers the content services to the users and possibly owns and manages the transportation network); *End Users (EU)* consumes the content; a *Content Provider (CP)* could exist, owning some *Content Servers (CS)*. However, DISEDAN does not deal with contractual CP-SP relationships; therefore one may assume that CSs are also owned by the SP. A solution is proposed for the (multi-criteria-hard) problem of best content source (server) selection, considering user context, servers' availability and requested content. The solution novelty consists in: (1) *two-step server selection mechanism* (at SP and at EU) using algorithms that consider context- and content-awareness and (2) *dual adaptation mechanism* consisting of *media adaptation* (also called *media flow adaptation*) and *content source adaptation* (by *streaming server switching*) when the quality observed by the user suffers degradation during the media session. The solution could be rapidly deployed in the market since it does not require complex architecture like Content Oriented Networking or CDNs [1] [2].

The *Dynamic Adaptive Streaming over Hypertext Transfer Protocol- HTTP (DASH)* technology has been selected for in-session media adaptation. The DASH was recently adopted as multimedia streaming standard, to deliver high quality multimedia content over the Internet, by using conventional HTTP Web servers [4] - [8]. It uses the HTTP protocol, minimizes server processing power and is video codec agnostic. Its basic concept is to enable automatic switching of quality levels according to network conditions, user requirements, and expectations. A DASH client continuously selects the highest possible video representation quality that ensures smooth play-out, in the current downloading conditions. This selection is performed on-the-fly, during video play-out, from a pre-defined discrete set of available video rates and with a pre-defined granularity (according to video segmentation). The DASH offers important advantages (over traditional push-based streaming), like: significant market adoption of HTTP and TCP/IP protocols to support the majority of the today Internet services; HTTP-based delivery avoids NAT and firewall-related issues; the HTTP-based (non-adaptive) deployment of progressive download existing today, can be conveniently be upgraded to support DASH; the ability to use standard/existing HTTP servers and caches instead of specialized streaming servers allows reuse of the existing infrastructure.

This work-in-progress is dedicated to take design decisions for a light Control Plane (CPl) and especially for its Monitoring subsystem (MON). The MON components are developed at SP, CS and optionally at EU Terminal (EUT). The MON is an essential DISEDAN component, contributing to the evaluation of the QoS and QoE. It is able to support both the initial server selection and then in-session actions.

Note that our main purpose here is not to essentially innovate in monitoring tools (a lot of implementations are available), but to integrate different components, aiming to develop a monitoring subsystem appropriate for DISEDAN light architecture.

Section II is a short overview of related work. Section III outlines the overall architecture and problem description. Section IV contains the paper main contributions, focused on defining CPl (Monitoring included) design decisions and implementation-related implications. Section V contains conclusions and future work outline.

 ISBN: 978-1-61208-411-4

II. RELATED WORK

The real-time adaptation in content streaming is a powerful and dynamic technique, adopted to solve the fluctuations in QoE/QoS. One can classify adaptation as acting on Media (flow) and/or on CS. The *Media adaptation* is a significant technique and main research innovation area in media streaming applications [6][7][13]. *CS adaptation* means a new content server selection (during the media session) and switching (handover), depending on the consumer device capabilities, consumer location, content servers state and/or network state [9][10].

A so-called "dual adaptation" is a process that integrates the above adaptation methods. The DISEDAN novel architecture [3] combines the initial server selection (result of cooperation between SP and EU) with session-time dual adaptation, in a single solution.

The initial server selection is based on optimization algorithms like *Multi-Criteria Decision Algorithms (MCDA)* [9][10], or *Evolutionary Multi-objective Optimization algorithm* (EMO) [11], modified to be applied to DISEDAN context. In these works several scenarios are proposed, analyzed and evaluated. In particular, the availability of different static and/or dynamic input parameters for optimization algorithms is considered. The result of this variability is that several CPl designs are possible, different in terms of performance and complexity. The dynamic capabilities for the initial CS selection and then for adaptation decisions depends essentially on the power of the DISEDAN monitoring system. It is the objective of this paper to analyze these variants, and define the monitoring subsystem.

The challenge in DISEDAN is to combine the DASH-related functionalities with additional monitoring in order to finally realize the dual adaptation.

The standard ISO/IEC 23009-1, "Information technology -- Dynamic adaptive streaming over HTTP (DASH)" [6], defines the DASH-Metrics client reference model, composed of *DASH access client (DAC)*, followed by the *DASH-enabled application (DAE)* and *Media Output* (MO) module. The DAC issues HTTP requests (for DASH data structures), and receives HTTP request responses. Consequently three observation points (interfaces – I/F) can be identified:

- O1 *at network-DAC I/F*: a set of TCP connections, each defined by its destination IP address, initiation, connect and close times; a sequence of transmitted HTTP requests, each defined by its transmission time, contents, and the TCP connection on which it is sent; and for each HTTP response, the reception time and contents of the response header and the reception time of each byte of the response body.

- *O2 at DAC-DAE I/F*: consists of encoded media samples. Each encoded media sample is defined as: media type; decoding time; presentation time; the *@id* of the Representation from which the sample is taken; the delivery time.

- *O3 at DAE-MO I/F*: consists of decoded media samples. Each decoded media sample is defined as: the media type; the presentation timestamp of the sample (media time); the actual presentation time of the sample (real time); the @id of the Representation from which the sample is taken (the highest dependency level if the sample was constructed from multiple Representations).

A summary of the metrics semantic defined in ISO/IEC 23009-1 [6], is: *Transmission Control Protocol (TCP) connections, HTTP request/response transactions, Representation switch events, Buffer level, Play list.* A similar list of QoE metrics standardized by 3GPP defined in 3GPP in 26.247, applicable for DASH, [8][13], contains: *HTTP request/ response transactions; Representation switch events; Average throughput; Initial play-out delay; Buffer level; Play list; MD information*.

III. DISEDAN SYSTEM ARCHITECTURE

A. General framework and assumptions

The definition and some details of the system architecture are already given in [3][9][10][12]. In this section, a summary only will be presented to support understanding of the CPl design decisions.

The main business entities/ actors are those mentioned in Section I: SP, EU, CS. The SP and CP entities are not seen as distinct in DISEDAN system. Also, a full CS management is out of scope of this system. The connectivity between CSs and EU Terminals (EUT) is assured by traditional *Internet Services Providers (ISP) / Network Providers (NP*) - operators. The ISP/NPs do not enter explicitly in the business relationships set considered by DISEDAN, neither in the management architecture (DISEDAN works in OTT style).

However, the DISEDAN solution can be also applied in more complex business models, e.g., involving Cloud Providers, CDN providers, etc. The relationships between SP and such entities could exist, but their realization is out of scope of this study. While Service Level Agreements (SLAs) might be agreed between SP and ISPs/NPs, related to connectivity services offered by the latter to SP, such SLAs are not directly visible at DISEDAN system level.

The system can work over the traditional TCP/IP mono and/or multi-domain network environment. The EUTs might not have explicit knowledge about the managed/non-managed characteristics of the connectivity services. No reservation for connectivity resources, neither connectivity services differentiation at network level are explicitly supposed (but they are not forbidden). This proves the system flexibility: it can work both in OTT style, or over a managed connectivity service offered by the network. Therefore, the SP does not commit to offer strong QoS guarantees for the streaming services provided to EUs. Consequently, DISEDAN does not suppose, but does not exclude, establishment of a SLA relationships between EUs and SPs management entities. However, it is assumed that a Media Description Server exists, managed by SP, to which EUT will directly interact.

The media streaming actions are independent on the transport networking technology. The EUT part (client side) works as a standalone client application, without any mandatory modifications applied to the SP; however, SP should provide some basic information to EUT, to help it in

 ISBN: 978-1-61208-411-4

making initial server selection (and optionally to help in-session CS switching). The decision about dual adaptation (media flow adaptation and/or CS switching will be taken mainly locally at EUT, thus assuring User independency and avoiding complex signaling between user and SP during the session.

Several CSs exist, known by SP (geographical location, server availability level, access conditions for users), among which the SP and/or EUs can operate servers selection and/or switching. The proposed architecture does not treat how to solve failures inside the networks, except attempts to do media flow DASH adaptation or CS switching.

The proposed system does not explicitly treat or innovate in the domain of content protection, Digital Rights Management (DRM), etc., but might use currently available solutions. Billing, financial aspects and other business related management of the DISEDAN high level services are out of the project scope.

The work [12], elaborated also in DISEDAN framework, has defined all requirements coming for EU, SP and out of them derived the general and specific System requirements, together with some assumptions and constrains imposed to such a system. The resulting high level architecture has been determined by such requirements. This work is based on the assumption of fulfillment of those requirements.

B. General Architecture

Figure 1 shows a simplified high level view of the general architecture.

The SP includes in its Control Plane:

- *MPD File generator* – dynamically generates Media Presentation Description (MPD) XML file, containing media segments information (video resolution, bit rates, etc.), ranked list of recommended CSs and, optionally - current CSs state information and network state (if applicable).
- *Selection algorithm* –runs Step 1 of server selection process. It exploits *MCDA* [9][10], modified to be applied to DISEDAN context, or *EMO* [11], etc., to rank recommended CSs and media representations, aiming to optimize servers load as well as to maximize system utilization.
- *Monitoring module* – collects monitoring information from CSs and performs the processing required to estimate the current state of each CS. Note that if some EU information should go to SP, then this information is transited (and aggregated) from EUT via CS towards SP.

The End User Terminal entity includes the modules:

- *Data Plane: DASH (access and application)* – parses the MD file received from SP and handles the download of media segments from CS; *Media Player* – playbacks the downloaded media segments.
- *Control Plane: Content Source Selection and Adaptation engine* –implements the dual adaptation mechanism; *Selection algorithm* –performs the Step 2 of server selection process. It can also exploit MCDA, EMO, or other algorithms to select the best CS from the set of candidates recommended by SP; *Monitoring module* – monitors changing (local) network and server conditions.

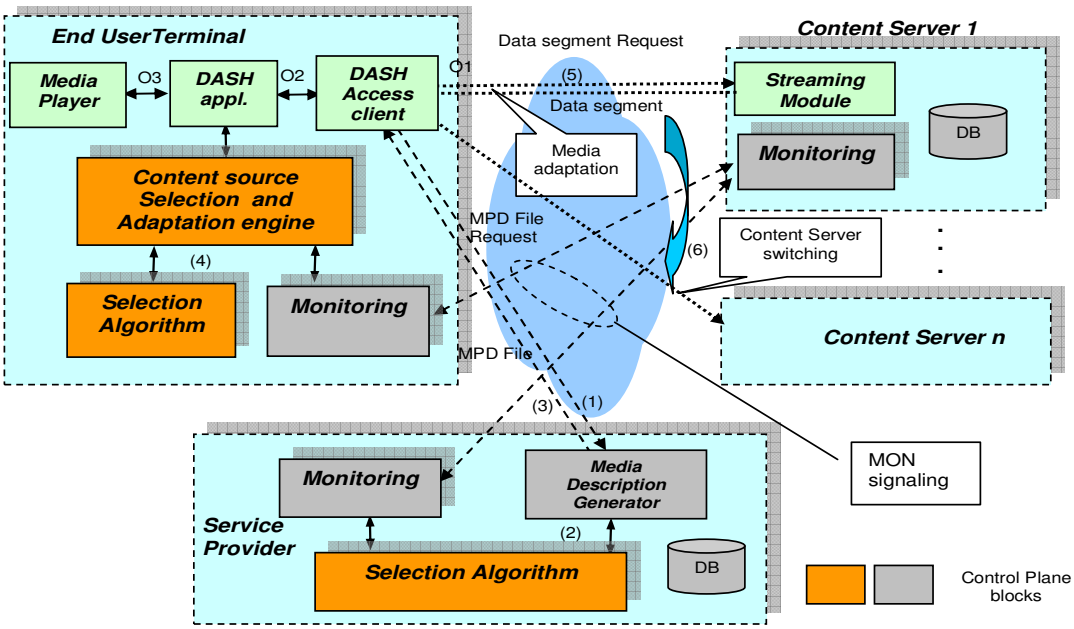


Figure 1. DISEDAN general architecture; DASH - Dynamic Adaptive Streaming over HTTP; MD – Media Description; DB – Data Base ; O1, O2, O3 – DASH Observation Points [ISO/IEC 23009-1]

The CS entity includes the modules:

- *Data Plane: Streaming module* – sends media segments requested by End Users; *Monitoring module* – monitors CS performance metrics (CPU utilization, network interfaces utilization, etc.). In a complex implementation of the CS, the monitoring can evolve from a simple probe to an advanced monitoring module, capable to supervise not only the active sessions but also some connectivity characteristics from this CS to different groups of users.

The following functional steps are performed (simplified description): (1) EUT issues to SP a media file request. (2) SP analyzes the status of the CSs and runs the selection algorithm (optionally the SP could make first, a current probing of the CSs); for each user request the SP could consider also the user profile, the policies of the SP for this user's class and other information at the SP side (e.g., state of the servers and possibly network-related information). (3) SP returns to EUT a ordered list of candidates CS (SP proposal) embedded in a MD- xml) file. (4) The EUT performs the final CS selection, by running its own selection algorithm and (5) starts asking segments from the selected CS. During media session the EUT makes quality and context measurements. Continuous media flow adaptation is applied using DASH technology, if necessary, or (6) CS switching is decided. From the EU point of view, the steps 1-2-3 composed the so-called Phase1 and steps 4-5-6 the Phase 2.

During the receipt of consecutive chunks, the user's application can automatically change the rate of the content stream (internal DASH actions- which are out of scope in this paper) and/or also can switch to another CS. When EU receives requested segments, it performs measurements to monitor parameters of download process. Note that the system is flexible in terms of monitoring procedures to follow. For instance if EU detects deterioration of downloading rate, it can use SP information about alternate CSs and/or it can start probing CSs. When the probing process finished, EU starts dual adaptation process to decide : media or server adaptation. If the first is selected, then EU downloads (via DASH) next segments with reduce rate, otherwise switches to another CS.

IV. MONITORING SUBSYSTEM

The architecture of the DISEDAN CPl is flexible. Several variants/versions of designs can be considered, i.e., a basic one or more complex, essentially depending on the roles of the business entities and their capabilities, interactions and also on SP and EU policies. The selection algorithms MCDA/EMO might work with different sets of static and/or dynamic input parameters. An important component of the CPl is the Monitoring *(MON@DISEDAN*).

A. Monitoring Architecture

Three MON modules have been identified in Figure 1: MON@SP, MON@CS, MON@EUT. However, not all these entities must participate to all phases of functioning. The variety of solutions determine the system overall performance but with additional cost for the more complex solutions. The monitored data are used to accomplish the following macro objectives :

- guide the initial server selection at SP and (optionally) at EU,
- guide the media adaptation and/or CS switching.

From the EUT point of view, two phases are distinguished: *Phase1* in which the EUT is not connected to any CS, but it just tries to do this, by contacting the SP; *Phase2* in which the EUT is currently served by a CS (media session time). The monitored data at EU level are different in Phase 1 w.r.t. Phase 2.

Note also that during media session, the DASH subsystem performs its own evaluation of the QoE and based on this, decides upon requested rate of the next video segment. The implementation of this type of monitoring is out of MON scope. However, the data collected from such on-line monitoring can be combined with other values delivered by MON@EU and delivered to other entities in the hierarchy (CS, SP). Actually, we adopted the approach described in [13] where it is recalled that in the 3GPP DASH specification TS 26.247 [7-8], QoE measurement and reporting capability is defined as an optional feature for client devices. If the EUT supports the QoE reporting feature, the DASH standard also mandates the reporting of all of the requested metrics at any given time; that is, the client should be capable of measuring and reporting all of the QoE metrics specified in the standard.

The standard TS 26.247 also specifies two options for the activation or triggering of QoE reporting: a. via the Quality Metrics element in the MPD; b. via the OMA Device Management (DM) QoE Management Object. In both cases a and b, the trigger message from the CS would include reporting configuration information such as the set of QoE metrics to be reported, the URIs for the server(s) to which the QoE reports should be sent, the format of the QoE reports, information on QoE reporting frequency and measurement interval, percentage of sessions for which QoE metrics will be reported, and access point names to be used for establishing the packet data protocol (PDP) context to be used for sending the QoE reports.

The selection algorithms MCDA/EMO might work with different sets of static and/or dynamic input parameters.

To achieve scalability of the monitoring system an important design decision is to avoid direct signaling between EUT and SP, except the initial request issued by EUT towards SP, in order to get the MPD xml file. Apart this phase, any monitored information obtained in EUT premises will be sent to the current CS serving that EUT.

We define three control bi-directional channels (see Figure 1) :

EUT-SP to generate the EU request to SP and to get the MD file from SP. This is performed in Phase 1 of the DISEDAN functional cycle, i.e., at CS selection time.

EUT–CS triggered by the serving CS, to report, the monitored data about current EU status and media session

 ISBN: 978-1-61208-411-4

data. This signaling is performed during Phase 2 time life for this EUT (i.e., media session).

CS-SP- to report: CS status data (capacity occupied, number of connections currently served, etc.); status data received from EUT (such data can be related to some individual users or aggregated at the CS level. The communication on this channel is triggered by the SP.

B. *Typical Scenarios*

Figure 2 presents a simplified *Message Sequence Chart (MSC)* illustrating the activities, communication in Data Plane (DASH) and the associated signaling executed in the Control Plane. One can see the Phase 1 and Phase 2 sets of actions, performed by EUT1.

Several types of monitoring activities are performed, described below.

Proactive monitoring: executed in some continuous mode (at SP level and possibly at EUT level- see the "loop" notations in Figure 2); such information is input for the CS selection algorithm (Phase 1), when some new content requests arrive from a given EU to SP. At SP, this means supervision of different servers, maybe networks, and user communities, depending on its policies. SP/CS cooperation on this purpose is envisaged. Such data can be also used to construct a history and updated status of the environment envisaged by the SP. The CSs could be involved in proactive monitoring, provided they are capable to probe the connectivity characteristics towards different groups of users (indicated by the SP).

At EU side, proactive monitoring might be performed, depending on capabilities of the EUT and its SW. In some more complex scenarios the EU can construct history, dedicated to its usual content connections (if they are estimated to be repeated in the future). The terminal context can be evaluated by such measurements, including its access network status.

In-session monitoring: monitoring is performed on a flow and data are collected in real time, to assess the level of QoS/QoE observed at EU side. These actions are basically performed by the EUT. Note that two kind of information are produced:

- collected by the DASH mechanisms, to serve internally as real time inputs to adaptation decision engine at EU,
- collected by the MON@EUT, which can be consolidated with those produced by the DASH, thus offering a more complete view not only about the reception of he media flow but also on general status and environment of the EUT.

In more complex DISEDAN variants, the SP and/or CS can be involved in such monitoring, at least in being aware of results (note that no SLA concerning mutual obligations of SP/EUs, related to QoE are established in DISEDAN system): for all active users or subsets; for all monitored data or summaries; full or summary monitored values.

Opportunity related monitoring: measurements essentially performed by the EUT to test the opportunity of switching the CS that delivers the content to EU. An example of such category is the Probing of some CS candidates if a CS switching action is prepared.

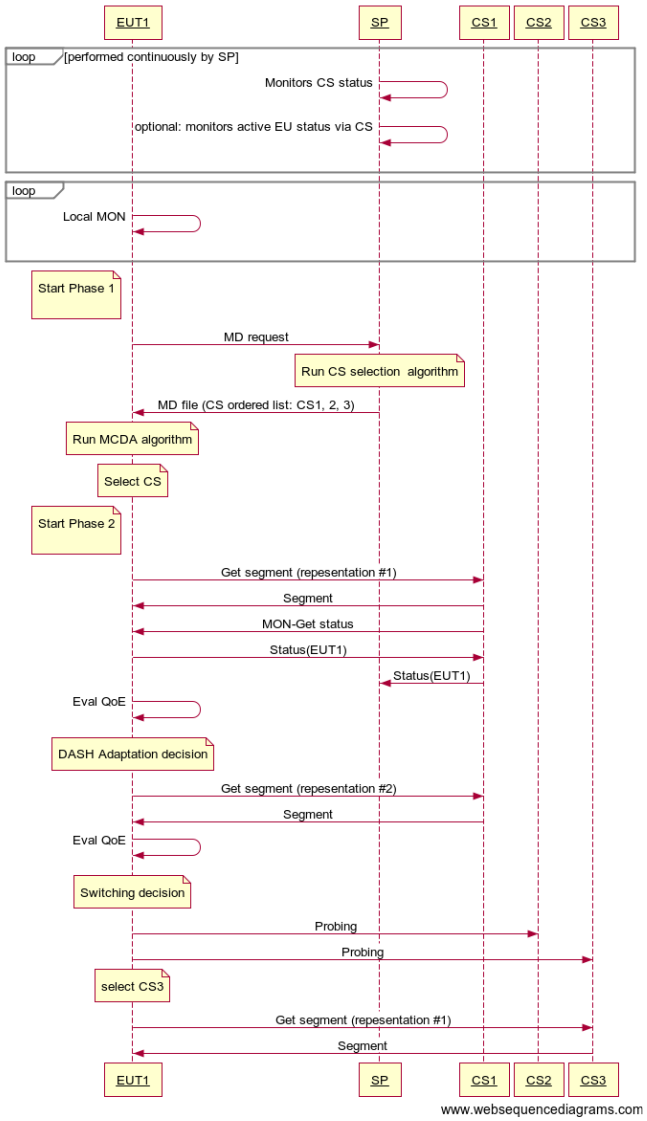


Figure 2. Typical activity and signaling diagram

C. *Metrics and MON versions*

Apart from DASH defined metrics (in-session observed), the MON subsystem may collect information on:

MON@EUT: CS accessibility (probing); EUT local dynamic context; historical and prediction data on servers and paths utilization.

*MON@SP:*CS status (collected from CS); active Users status; current load on some paths (here the network monitoring of the NP should cooperate); other dynamic, characteristics of some paths (e.g., loss, jitter); historical and prediction data on servers and paths utilization.

*MON@CS:*CS status (load); CS environment data (network paths, connectivity paths dynamic characteristics - evaluated at overlay level - from CS to different groups of users; EUTs data, active user groups data.

 ISBN: 978-1-61208-411-4

Therefore the overall MON system design is flexible, since it can combine different features of the above components.

D. Implementation aspects

SP and CS will have an internal database that will contain monitored and/or post-processed data. Also these two entities will be capable to send and receive JSON messages embedded in simple HTTP calls. EU might not have any internal database; it will just have the basic capability to send only simple HTTP calls to either SP or CS.

For Database is proposed to use *PostgreSQL*, technology [16]. The *PostgreSQL* is a powerful, open source object-relational database system. It runs on all major operating systems, including Linux, UNIX, and Windows.

SP and CS must be able to receive and send simple HTTP messages to each other. For this reason it is needed need a web server and a programming language to implement these features. Web server of choice is Node.js [17].

The *Node.js* [17] is an open source, cross-platform runtime environment for server-side and networking applications. The *Node.js* applications are written in JavaScript, and can be run within the *Node.js* runtime on OS X, Microsoft Windows, Linux and FreeBSD.

Currently the system is under implementation phase, performed by the DISEDAN consortium. A pilot system has been constructed having a core network (three IP network domains, independently managed, a SP entity and several CSs and EUTs distributed onto the local area networks linked to the core. Preliminary tests showed that MCDA applied at server level [10][12], produces the best trade-off selection of the server-path pair offering a good QoE to the EU in rather loaded network conditions.

V. CONCLUSIONS AND FUTURE WORK

This paper presented the design concepts and decisions for the CPl of a *media delivery system having a light-architecture and working on top of the current Internet connectivity.*

The work focus is on the Monitoring subsystem, seen as a main component to provide dynamic information, to support the two major functional phases: initial CS selection and then in-session actions for media adaptation and/or CS switching. The architectural specification and then the design of the monitoring system have been proposed, combining the DASH - embedded monitoring features (to evaluate QoS/QOE) with external-to -DASH monitoring functions, thus completing the updated image of the DISEDAN environment (End User, Content Servers, network).

Future experimental results of the implementation will be reported in another paper.

ACKNOWLEDGMENT

This work has been partially supported by the Research Project DISEDAN, No.3-CHIST-ERA C3N, 2014- 2015.

REFERENCES

[1] J. Choi, J. Han, E. Cho, T. Kwon, and Y. Choi, “A Survey on Content-Oriented Networking for Efficient Content Delivery”, IEEE Communications Magazine, March 2011, pp. 121-127.

[2] ***, “Information-Centric Networking-3”, Dagstuhl Seminar, July 13-16 2014, http://www.dagstuhl.de/en/program/calendar/semhp/?seminar=14291, [retrieved: February, 2015].

[3] *** http://wp2.tele.pw.edu.pl/disedan/ [retrieved:May, 2015]

[4] T. Dreier, "Netflix sees cost savings in MPEG DASH adoption," 15 December 2011. [Online]. Available: http://www.streamingmedia.com/Articles/ReadArticle.aspx?ArticleID=79409. [retrieved: October, 2014].

[5] I. Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet," MultiMedia, IEEE, vol. 18, no. 4, 2011, pp. 62 - 67.

[6] ISO/IEC 23009-1, "Information technology -- Dynamic adaptive streaming over HTTP (DASH) -- Part 1: Media presentation description and segment formats," ISO/IEC, Geneve, second edition, 2014.

[7] 3GPP TS 26.247 v10.1.0, “Transparent End-to-End Packet Switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming Over HTTP (3GP-DASH),” Release 10, June 2011.

[8] ETSI TS 126 247 V11.7.0 (2014-07) (UMTS); LTE; Transparent end-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH), (3GPP TS 26.247 version 11.7.0 Release 11, 2014).

[9] A. Beben, J. M. Batalla, W. Chai and J. Sliwinski, “Multi-criteria decision algorithms for efficient content delivery in content networks”, Annals of Telecommunications - annales des telecommunications, Springer, vol. 68, Issue 3, 2013, pp. 153-165.

[10] E. Borcoci, M. Vochin, M. Constantinescu, J. M. Batalla, and D. Negru, “On Server and Path Selection Algorithms and Policies in a light Content-Aware Networking Architecture”, ICSNC 2014 Conference, http://www.elcom.pub.ro/disedan/docs/ICSNC%202014%20Conf.pdf.

[11] J. M. Batalla, C.X. Mavromoustakis, G. Mastorakis, D. Négru and E. Borcoci, “Evolutionary Multiobjective Optimization algorithm for two-phase content source selection process in Content Aware Networks”, submitted to Springer 4OR - A Quarterly Journal of Operations Research.

[12] E. Borcoci, ed., et al., D2.1 “System requirements and comparative analysis of existing solutions for media content server selection and media adaptation”, July 2014, http://wp2.tele.pw.edu.pl/disedan .

[13] O. Oyman and S. Singh,"Quality of Experience for HTTP Adaptive Streaming Services", IEEE Communications Magazine, April 2012, pp.20-27.

[14] C. Alberti, D. Renzi, C. Timmerer, C. Mueller, S. Lederer, S. Battista and M. Mattavelli, “Automated QoE Evaluation of DASH”, 2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX), http://infoscience.epfl.ch/record/188813/files/p20-Alberti.pdf, [retrieved: February, 2015].

[15] C. Müller, S. Lederer and C. Timmerer, “An Evaluation of Dynamic Adaptive Streaming over HTTP in Vehicular Environments” , www-itec.uni-klu.ac.at/dash/?p=517.

[16] ***, http://www.postgresql.org, [retrieved: March, 2015].

[17] ***, http://nodejs.org, [retrieved: March, 2015].

 ISBN: 978-1-61208-411-4

Mean Opinion Score Measurements Based on E-Model During a VoIP call

A Single Comparison On-line and Off-line

Edgard Silva, Leandro Galvão, Edjair Mota
Institute of Computing – ICOMP
Federal University of Amazonas – UFAM
Manaus – Brazil
e-mail: {edgard, galvao, edjair}@icomp.ufam.edu.br

Yuzo Iano
Department of Communications – DECOM
University of Campinas – UNICAMP
Campinas – Brazil
e-mail: yuzo@decom.fee.unicamp.br

***Abstract*— This paper presents the Mean Opinion Score (MOS) measurements results of an input stage of an adaptive speech based on Quality of Experience (QoE) control during a Voice Over Internet Protocol (VoIP) call. QoE is periodically determined by means of a modified version of E-model and recorded in a log file. At the end, we compared the results to POLQA (Perceptual Objective Listening Quality Assessment), and prove that continuously measurements are correct and can be used in the future real-time QoE controller to be build.**

Keywords-Quality of Experience (QoE); Voice over Internet Protocol (VoIP); Adaptive control system; MOS; speech quality; Codec Switching.

I. INTRODUCTION

Present in our personal and professional activities, speech signals transmissions over computer networks (VoIP - Voice over Internet Protocol) have gained wide acceptance by general Internet users. In a VoIP application, the main goal is to transfer voice signals over the IP network. In order to achieve this, the voice is digitized and packetized at the sender, and next, the result of the packaging is transmitted over the IP network to the receiver. In the receiver, the voice signals are unpacked from the received packets and then are decoded. After this, the results are played out to the listener [1]. However, VoIP is subject to several degradations, both at the application layer or the network layer, such as compression of the encoder, the delay end-to-end packet loss, jitter and bandwidth levels. Thus, to keep and attract new users, the quality of the provision of VoIP services need to be measured and optimized to ensure user satisfaction [2].

Quality provision assurance is one of the problems not solved yet, although VoIP enjoys the progress made in the last two decades [3]. In recent years, researchers developed Quality of Service (QoS) Control mechanisms to improve the use of network resources and user's terminal to minimize speech quality degradation. Some of these mechanisms seek to adapt the voice stream or other VoIP parameters, according to significant changes in network end users preferences, or providing requirements of service providers.

The speech quality can be measured subjectively or objectively. Subjective evaluation involves 12-24 participants individually listening to an audio stream of several seconds and classifying the audio quality on a scale of 1 (poor) to 5 (excellent). These ratings form a single Mean Opinion Score (MOS), as specified in ITU-T Recommendation P.800 [4]. This evaluation is costly and time-consuming and it cannot be done in real-time if one considers these characteristics. Thus, various techniques have been proposed to estimate objectively MOS (without human perception), such as POLQA [5] and E-model [6].

POLQA [5] is a perceptual technique that compares off-line two signals to generate the MOS: a reference signal (for example, captured at the sender) and the degraded signal (for example, captured at the receiver). The requirement makes the approach unsuitable for live call monitoring. On the other hand, the technique of E-model specified in ITU-T Rec. G.107 [6] is a noninvasive method that uses the network metrics monitored locally and the equipment impairment factor to estimate the quality of the call, so it can be used for monitoring live calls. A problem with the E-model is that only the ITU-T provides the equipment impairment factor specified in ITU-T Rec. G.113 [7]. For a range of other commonly used codecs not specified by the ITU-T Rec. G.113 [7], the equipment impairment factor is not provided.

Adaptive control systems in general respond to changes in their internal state or external environment with guidance of an underlying control system. VoIP systems are likely to need dynamic adaptation to deal with the complex dilemma between voice quality and impairment. This is necessary due to the nature of decentralized control of IP networks and the stochastic nature of data packets delivery. While existing solutions for QoS control of VoIP show some performance improvement and have feedback, they do not provide explicit focus on the control loop [8]. Measurable QoS relates directly to the state of the network, while QoE relates directly the quality level perception that users have. This perception will play a key role in the decision of making a VoIP application success or failure [9].

In this paper, our intention is to obtain measurements in real-time, of QoE of a VoIP transmission. In the future, our intention is be to develop a robust controller for a codec in command of a bidirectional audio stream.

The outline of the paper is as follows. The next section briefly surveys the issue concerning the adaptation during a VoIP call. Section III presents the methodology we followed to get measures of the metrics of interest. This section also presents details about the scenarios for testing and the measurements results while Section IV discusses the findings. A brief concluding section presents our ongoing research.

 ISBN: 978-1-61208-411-4

II. BACKGROUND

The studies conducted by Karapantazis [3], Manousos [10], Costa [11], Qiao [12], Myakotnykh [13] and Viana [14] show that adaptation during a VoIP call (e.g., codec, packing, redundancy) can significantly improve the speech quality. However, as pointed out by Carvalho [15], these works often focus on parameter setting and little or nothing are based on the advances of research in adaptive systems.

Aktas et al. [16] compares the speech quality of a set of standard VoIP codecs given different network conditions and propose an adaptive end-to-end based codec switching scheme based on packet loss, jitter, and available bandwidth as the factors that define the current network condition.

Costa and Nunes [17] describe an adaptive codec switching technique that starts to monitor and analyze the quality of the voice, changing to a lower or higher codec rate according to predefined threshold values for each codec.

Haytham Assem [1] presents and evaluates an algorithm that performs the selection of the most appropriate audio codec given prevailing conditions on the network path between the endpoints of a voice call.

Bringing together contributions from the fields of VoIP and adaptive control systems, this project proposes a well-founded solution to the problem of real-time control of the quality of speech on VoIP calls and introduces the method of measurement and results the input signal of the proposed controller. For this, it is necessary to define what to measure and how to measure these parameters of the controller input.

III. METHODOLOGY

The method to be employed in this work is to improve the diagnostic module developed in [15]. The diagnostic module is composed of agents of monitoring and analysis. It uses the RTCP XR reports (RTP Control Protocol Extended Reports) [18]. These reports carry information about the instantaneous quality of a VoIP call, such as loss, delay and ambient noise, codec impairment, all this used in a function that calculates the E-model for instant call quality.

As a testbed, we used 64-bit machines with Ubuntu 10.04 operating system (newer versions proved inefficient for preliminary tests and incompatible with the Intel IPP 7.7.1, a codec library used in the experiments) to implement the PJSIP 1.10, a free library of multimedia communication and open code written in the C language that implements the protocols based on standards such as SIP, SDP, RTP, among others [19].

We had in the first instance, a machine (sender) directly connected to another machine (receiver) as proof of concept for local testing and controlled conditions.

At all stages of this experiment, we recorded logs files of transmission (sent file) and received (received file) files for analysis in Section VII. We transmitted the same file multiple times in different test scenarios and test multiple files with different contents in order to compare the measurements of the systems (internal measurements) and the result of the global measurements (POLQA).

We have two kinds of moments of measurements: during a VoIP call (punctual) and the cumulative average at the end of a VoIP call (global).

This paper presents the results of measurements performed during a call, the procedure and the results graphically.

During a VoIP call, there are two applications for point measurements, which are results of the measurements provided by the E-model routine implemented in PJSIP via RTCP-XR:

- Communication Performance: the instantaneous value of the QoE during an analysis interval of a VoIP call.
- Control: (Specifically in this study, the control application is not enabled).

We have the E-model (Rec. ITU-T G.107 [6]) among the methods of timely measurement of quality of speech most used. Its measurement procedure consists of collecting parameters of voice stream, which serve as input to a set of equations that return as a result the R factor, whose value ranges from 0 (worst) to 100 (best) as a measure of quality speech evaluated. This result is mapping from R factor (0-100) to MOS (1-5) by (1).

$$MOS = \begin{cases} 1 & if\ R < 0 \\ 1 + 0.035R + 7\text{x}10^{-6}R & if\ 0 \le R \le 100 \\ (R-60)(100-R) & \\ 4.5 & if\ R > 100 \end{cases} \quad (1)$$

Typically, the result of E-model is transformed to a scale of MOS [4] as presented in Table I.

At the end of the VoIP call, we received a file (received file) for a given file transmitted (sent file) in a given situation, to be used for measuring the overall quality of the call during the analysis presented in Section IV, and a score between 1 and 5 is generated, including a confidence interval.

IV. ANALYSIS OF RESULTS

Having two types of metrics (punctual/local and global/external) we had two stages of analysis.

Quality measurement based on QoE will be held along the VoIP call. The measurement data quality along the VoIP call will be determined by a series of calls, with the same uploaded file (sent file), with and without the controller application enabled. Then, the results will be compared each other. A graphic of the call quality over time will be generated (in seconds) (Q × t) for both situations, considering the uncertainty range for each measurement point in the various transmissions.

The signal received by the listener (received file) will be recorded and compared off-line with the transmitted signal (sent file) with a perceptual objective method of measuring quality of speech and should be as close as possible to the subjective quality scores obtained in subjective tests hearing.

 ISBN: 978-1-61208-411-4

TABLE I. MEAN OPINION SCORE VALUE [16]

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

This method uses the knowledge of the workings of the human auditory system to compare a reference signal with a degraded signal in order to compose a measure of voice distortion, the main representatives sit POLQA, also known as ITU-T Rec. P.863 [5], which includes a model for predicting voice quality through analysis of digital voice signal.

The data collected will be analyzed by statistical tools like Akaroa [20] and R [21], based on statistical hypothesis testing, longitudinal data analysis, among other techniques. In order to emulate the packet loss in our network scenario, we use the Netem (Network Emulation) [23].

In this work, voice files from the Open Speech Repository (OSR) were used [22]. Figures 1, 2 and 3 are related to the sent file, named osr_us_000_0010_8k.wav, a female voice file, 16-bit PCM, 8 kHz sample rate. The sender establishes a VoIP call to the receiver machine and sends this file. A total of 25 connections were established for the same file at the following network conditions controlled by Netem: 0%, 2.5%, 5%, 7.5% and 10% packet loss. In all cases, the G.711 codec was used. The log files generated by each transmission feed a Perl script that filters the instantaneous quality information. An R script determines and plots the average value of instantaneous measured signal quality. This average is compared to the result of the analysis with POLQA of received and transmitted audio files.

Instantaneous measurements of MOS quality were taken during the random interval of 0.5 to 1.5 seconds. As the packet loss increases, the instantaneous MOS quality has become more unstable, as presented in Figure 1. This occurs as the RTP packets of data travelling across a network fail to reach their destination, i.e., the receiver station.

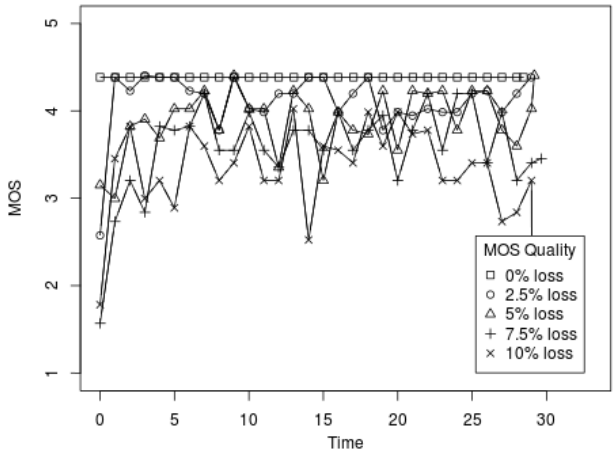


Figure 1. Instantaneous MOS Quality

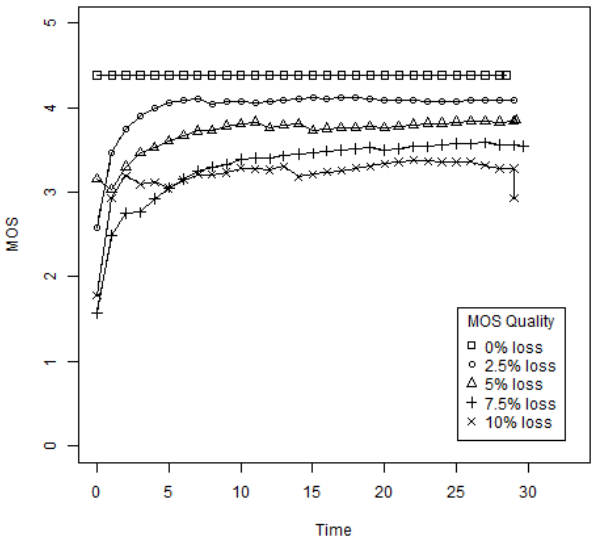


Figure 2. Average MOS Quality

Figures 1, 2 and 3 show the MOS in time considering some impairment. They show that the impairments result in MOS reduction.

As the value of MOS is calculated from the R factor and this is directly related to the network parameters, one being the loss of packets, by varying the value of the R factor, the instantaneous value of the MOS also varies. A variation on the R factor necessarily implies a variation of the MOS. Note that the best case for the MOS is when there is no packet loss.

The average value of the MOS Quality approaches to the value of the global MOS as time passes. In Figure 2, the MOS value for a given situation of packet loss gets worse as the loss increases.

In Figure 2, the average MOS Quality presents a little difference before 5 seconds between the values of packet loss of 5% to 10%. The initial values have wide oscillations that are minimized by the averaging as the time elapses.

As the packed loss increases, MOS Quality decreases, as shown in Figure 3.

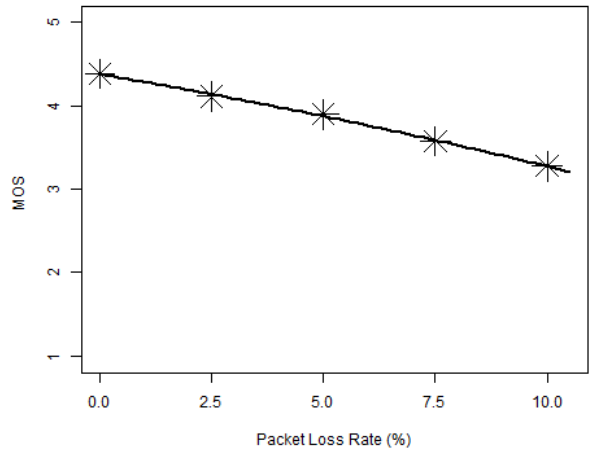


Figure 3. Average MOS versus packed loss rate - measured approximated values

V. CONCLUSIONS AND FUTURE WORK

This paper presented the results of measurements of input parameters of an adaptive control system of quality of speech based on QoE during a VoIP call.

The QoE varies along a voice transmission depending on factors like packet loss, jitter and bandwidth in addition to the equipment impairment factor. It is possible to measure these values and transform them into a single value (R factor). The R factor and the value of MOS Quality can be used as input variables in our adaptive control system. They vary over time and its measurement was easily implemented. This will lead our control system to make a decision to change or not the encoder and when this change will be made in order to impact as little as possible the communication. The current encoder itself represents one of the input parameters in the control system; all of that seeks to minimize the encoder changes in order to maintain the best quality experience for the users.

The system will choose a new codec according to the trend of monitored variables. Machine learning can be used as new encoders appear on the market. Three factors must be considered: a) when to switch the codecs, b) the codec used and the codec chosen to replace it and c) the reasons that led to the decision by the choice of the new codec.

The next steps of this work will be the extent of equipment impairment factor for no ITU codecs, such as Speexs, for example. Moreover, add to library encoders available, the Opus [8][24] wideband codec.

ACKNOWLEDGMENT

We would like to thank the School of Technology (EST) of the State University of Amazonas (UEA), the Foundation for Research Support of the State of Amazonas (FAPEAM), the Department of Communications of the Faculty of Electrical and Computer Engineering of the University of Campinas (UNICAMP), FAPESP, CAPES and CNPq for support in developing this work.

REFERENCES

[1] A. Haytham, M. Adel, D. Malone, B. Jennings, J. Dunne, and P. O'Sullivan, "A Generic Algorithm for Mid-Call Audio Codec Switching," in *IFIP/IEEE International Workshop in Quality of Experience Centric Management (QCMAN 2013)*, Ghent, Belgium, May 2013.

[2] D. Rodrigues, E. Cerqueira, and E. Monteiro, "QoE Assessment of VoIP in Next Generation Networks," in *Proceedings of the 12th IFIP/IEEE International Conference on Management of Multimedia and Mobile Networks and Services: Wired-Wireless Multimedia Networks and Services Management*, Venice, Italy, October, 2009.

[3] S. Karapantazis and F.-N. Pavlidou, "VoIP: A comprehensive survey on a promising technology," *Computer Networks,* vol. 53, 2009, pp. 2050–2090.

[4] ITU-T, "Rec. P.800: Methods for subjective determination of transmission quality," Geneva (Switzerland), 08/1996.

[5] ITU-T, "Rec. P.863: Perceptual objective listening quality assessment (POLQA)," Geneva (Switzerland), 01/2011.

[6] ITU-T, "Rec. G.107: The E-model, a computational model for use in transmission planning," Geneva (Switzerland), 07/2002.

[7] ITU-T, "Rec. G.113: Transmission impairments due to speech processing Appendix I: Provisional planning values for the equipment impairment factor Ie," Geneva (Switzerland), 10/2001.

[8] J.-M. Valin, K. Vos, and T. B. Terriberry, "Definition of the Opus Audio Codec, RFC6716," September 2012. [Online]. Available: http://www.rfc-base.org/txt/rfc-6716.txt. [retrieved: Feb. 2014].

[9] M.-D. Cano and F. Cerdan, "Subjective QoE analysis of VoIP applications in a wireless campus environment," *Telecommunication Systems,* vol. 49, 01 January 2012, pp. 5-15.

[10] M. Manousos, S. Apostolacos, I. Grammatikakis, D. Mexis, D. Kagklis, and E. Sykas, "Voice-quality monitoring and control for VoIP," *Internet Computing,* vol. 9, no. 4, July-Aug 2005, pp. 35-42.

[11] N. Costa and M. S. Nunes, "Dynamic Adaptation of Quality of Service for VoIP Communications," vol. 2, no. 2 and 3, 2009, pp. 155-166.

[12] Z. Qiao, L. Sun, N. Heilemann, and E. Ifeachor, "A new method for VoIP quality of service control use combined adaptive senter rate and priority marking," in *IEEE International Conference on Communications*, 2004.

[13] E. S. Myakotnykh and R. A. Thompson, "Adaptive Rate Voice over IP Quality Management Algorithm," *International Journal on Advances in Telecommunications,* vol. 2, 2009, pp. 98-110.

[14] B. d. A. Vianna, N. T. Moura, C. V. N. d. Albuquerque, V. E. F. Rebello, and C. Boeres, "adaMOS: Algoritmo MOS-Adaptativo para Fontes VoIP," in *WebMedia '06 Proceedings of the 12th Brazilian Symposium on Multimedia and the web*, New York, NY, USA, 2006.

[15] L. S. G. d. Carvalho, "Adaptive Management of Speech Quality Between VoIP Terminals," PhD thesis, Universidade Federal do Amazonas (UFAM), Manaus, 2011.

[16] I. Aktas, F. Schmidt, E. Weingärtner, C.-J. Schnelke, and K. Wehrle, "An Adaptive Codec Switching Scheme for SIP-Based VoIP," in *12th International Conference, NEW2AN 2012, and 5th Conference, ruSMART 2012*, St. Petersburg, Russia, Jan. 2012.

[17] N. Costa and M. S. Nunes, "Adaptive Quality of Service in Voice over IP Communications," *Proceedings of the 2009 Fifth International Conference on Networking and Services (ICNS 2009)*, Abril 2009, pp. 19-24.

[18] T. Friedman, R. Caceres, and A. D. Clark, "RTP Control Protocol Extended Reports (RTCP XR). Request for Comments (RFC) 3611," November 2003. [Online]. Available: https://tools.ietf.org/html/rfc3611.

[19] B. Prijono and P. Ismangil, "PJSIP," [Online]. Available: http://www.pjsip.org/. [retrieved: Feb. 2015].

[20] G. Ewing, K. Pawlikowski, and D. McNickle, "Akaroa-2: Exploiting network computing by distributing stochastic simulation," in *Proceedings of the 13th European Simulation Multi-conference*, Warsaw, Poland, 06/1999.

[21] R Core Team, "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, 2013. [Online]. Available: http://www.R-project.org. [retrieved: Feb. 2015].

[22] O. Project, "Open Speech Repository," [Online]. Available: http://www.voiptroubleshooter.com/open_speech/american.html. [retrieved: Feb. 2015].

[23] Linux Foundation, "Netem: Network Emulator," 19 November 2009. [Online]. Available: http://www.linuxfoundation.org/collaborate/workgroups/networking/netem. [retrieved: Feb. 2015].

[24] IETF, "Definition of the Opus Audio Codec, RFC6716," IETF, 09 2012. [Online]. Available: http://www.rfc-base.org/txt/rfc-6716.txt. [retrieved: Feb. 2015].

[25] H. Müller, M. Pezzè, and M. Shaw, "Visibility of control in adaptive systems," in *Proceeding ULSSIS '08 Proceedings of the 2nd international workshop on Ultra-large-scale software-intensive systems*, New York, NY, USA, 2008.

 ISBN: 978-1-61208-411-4

QoE-Based Adaptive Control of Speech Quality in a VoIP Call

Edgard Silva, Leandro Galvão, Edjair Mota
Institute of Computing – ICOMP
Federal University of Amazonas – UFAM
Manaus – Brazil
e-mail: {edgard, galvao, edjair}@icomp.ufam.edu.br

Yuzo Iano
Department of Communications – DECOM
University of Campinas – UNICAMP
Campinas – Brazil
e-mail: yuzo@decom.fee.unicamp.br

***Abstract*—This study presents the idea of building a robust controller for a codec in charge of a two-way audio stream bringing together the contributions of the areas of Voice over IP (VoIP) and adaptive systems. Such a controller will be possible with the application of a solution based on the Quality of Experience (QoE) for the control problem real-time speech quality of VoIP calls for the future applications and services.**

Keywords-Quality of Experience (QoE); Voice over Internet Protocol (VoIP); Adaptive control system; speech quality; Codec Switching.

I. INTRODUCTION

Speech transmission over computer networks is liable to several impairments from both application and network layer, such as codec compression, end-to-end delay, and packet loss. In the last years, Quality of Service (QoS) control mechanisms have been developed to make optimum use of network and terminal resources in order to minimize the effects of network impairments on speech quality. Some of these mechanisms seek to adapt the voice flow or other VoIP-related parameters in accordance with significant changes in the network, end user's preferences, or service provider's requirements.

Adaptive systems in general respond to changes in their internal state or external environment with the guidance of an underlying control system. VoIP systems are particularly likely to require a dynamic adaptation solution for dealing with the complex trade-off between speech quality and impairments, due to the decentralized control nature of IP networks and the stochastic nature of data packet delivery. Although the existing adaptive solutions for QoS control of VoIP show some performance improvement and exhibit some feedback, they do not provide explicit focus on the control loop [1].

In this paper, we aim to develop a robust controller for a codec in charge of a bidirectional audio flow. This controller will take some observation variables as input, such as latency and packet loss, and map them into adjustable variables, such as packetization, bitrates, sampling frequency, redundancy level, among others. The control objective will be initially set towards speech quality performance, in terms of QoE. It can also be extended to other issues, such as energy consumption, resource optimization, security aspects, and so on.

Figure 1 and Figure 2 give an overview of the controller to be developed. Figure 1 shows the feedback loop that lies at the core of any self-adaptive system. The feedback loop, also known as adaptation or autonomic loop, typically involves four key activities: monitoring, analysis, planning, and execution [2][3].

As depicted in Figure 1, sensors collect data from the managed system. The feedback cycle starts with the monitoring of relevant data that reflect the current state of the system. Next, the system analyzes the collected data, structuring and reasoning about the raw data. Upon completing this step, decisions must be planned about how to adapt the system to reach a desirable state. Finally, to implement the decision, the system must execute it by means of available effectors. Central to this loop, there will be a knowledge base that keeps the necessary information about the managed entities and their operations.

Current VoIP solutions for QoS control of speech quality lack of this view. Bringing the control loop to surface can improve the efficiency of such solutions.

Whereas Figure 1 shows the agents that compose the control loop, Figure 2 shows the information flow among these agents. Usually, a system converts input signals into output signals by performing operations on the inputs and intermediate products. The values of measurable properties of system's states are called variables [4]. A first step in designing an adaptive mechanism is to identify the key variables of the managed system:

- Observation parameters. They are measurable variables from which the adaptive mechanism can infer the status of the managed system.
- Decision metrics. They characterize the system performance over a sampling period and that the planning agent tries to optimize. They can be equivalent to a single observation parameter, such as delay and packet loss, or a synthesis of a set of observation parameters, such as Mean Opinion Score (MOS).
- Performance references. They represent the desired system performance in terms of observation parameters.
- Adjustable parameters. They correspond to the effectors in the feedback loop (Figure 1), an attribute of the managed system that can be manipulated to apply the necessary adaptations.

Essentially, adaptive systems implement a transfer function that takes decision metrics as input and gives the amount of change (if needed) in the adjustable parameters as output.

 ISBN: 978-1-61208-411-4

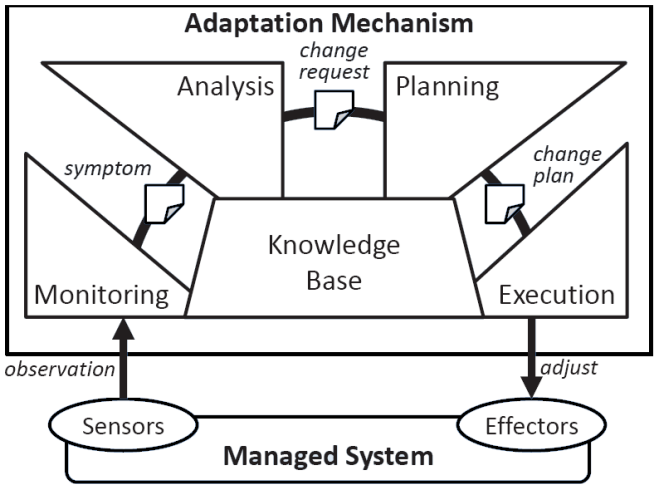


Figure 1. The feedback loop.

This project will be a follow-up of the work of Carvalho [5], in the sense of exploring newer codecs, such as Opus [6], and making use of a larger set of adjustable parameters in order to adapt the audio flow to monitored network conditions.

A. Experimental Techniques

The four classes of agents of the feedback control loop (Monitoring, Analysis, Planning and Execution) in Figure 1 can be arranged in different ways in order to control an audio flow between two endpoints (sender and receiver). Some of these arrangements will be implemented as candidate controllers. Hence, some comparative tests will be performed to select the best arrangement of the control loop agents.

The validation of the adaptive controller will be based on measurements over both simulated and real Internet scenarios. The audio flow will be systematically submitted to some network impairments, and its performance will be measured in terms of latency, packet loss, and MOS. Those measurements will support the researcher to check the candidate codec controllers against the following self-adaptive properties [7]: stability, accuracy, short settling time, small overshoot, robustness, and scalability.

The collected data will be analyzed by statistical tools like Akaroa [8][9] and R [10], based on statistical hypothesis testing, longitudinal data analysis, among other techniques.

B. Work organization

The project execution is divided into two basic steps: controller implementation and experimentation. During the controller implementation step, the researchers will design and implement an adaptive controller for an audio codec. During the experimentation phase, the controller candidates will have their performance compared in order to determine which arrangement of the control loop agents is more robust against network impairments.

The software development steps will be conducted by a PhD candidate and a student. The PhD candidate will specify the audio flow parameters that should be monitored and adjusted by the controller, and the arrangements of control loop agents that will be tested.

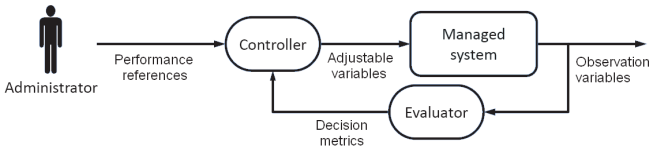


Figure 2. Variables related to a generic adaptive system.

Also, he or she will plan the experimental setup, conduct the statistical analysis of the collected data and look for improvements in controller design, code writing or experimental procedures.

The student will be responsible for writing the code of the controller candidates, and executing and automating the experimental apparatus.

ACKNOWLEDGMENT

We would like to thank the School of Technology (EST) of the State University of Amazonas (UEA), the Foundation for Research Support of the State of Amazonas (FAPEAM), the Department of Communications of the Faculty of Electrical and Computer Engineering of the University of Campinas (UNICAMP), FAPESP, CAPES and CNPq for support in developing this work.

REFERENCES

[1] H. Müller, M. Pezzè, and M. Shaw, "Visibility of control in adaptive systems," em *Proceeding ULSSIS '08 Proceedings of the 2nd international workshop on Ultra-large-scale software-intensive systems*, New York, NY, USA, 2008.

[2] H. A. Müller, H. M. Kienle, and U. Stege, "Autonomic Computing Now You See It, Now You Don't," em *Software Engineering International Summer Schools, ISSSE 2006-2008, Revised Tutorial Lectures*, vol. 5413, Salerno, Springer Berlin Heidelberg, 2009, pp. 32-54.

[3] M. C. Huebscher and J. A. McCann, "A survey of autonomic computing—degrees, models, and applications," *ACM Computing Surveys (CSUR)*, vol. 40, n. 3, August 2008, pp. 7.1-7.28.

[4] M. Shaw, "Beyond objects: a software design paradigm based on process control," *ACM SIGSOFT Software Engineering Notes*, vol. 20, n. 1, January 1995, pp. 27-38.

[5] L. S. G. Carvalho, "Adaptive Management of Speech Quality Between VoIP Terminals," PhD thesis, Universidade Federal do Amazonas (UFAM), Manaus, 2011.

[6] J.-M. Valin, K. Vos, and T. B. Terriberry, "Definition of the Opus Audio Codec, RFC6716," September 2012. [Online]. Available: http://www.rfc-base.org/txt/rfc-6716.txt. [Retrieved: Feb. 2015].

[7] N. M. Villegas, A. H. Müller, G. Tamura, L. Duchien, and R. Casallas, "A Framework for Evaluating Quality-Driven Self-Adaptive Software Systems," em *SEAMS '11 Proceedings of the 6th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, Waikiki, Honolulu, HI, USA, May, 2011.

[8] G. Ewing, K. Pawlikowski and D. McNickle, "Akaroa-2: Exploiting network computing by distributing stochastic simulation," em *Proceedings of the 13th European Simulation Multi-conference*, Warsaw, Poland, 06/1999.

[9] E. S. Mota, A. Wolisz, and K. Pawlikowski, "A perspective of batching methods in a simulation environment of multiple replications in parallel," em *Simulation Conference*, Orlando, FL, USA, 10 Dec 2000-13 Dec 2000, vol.1, no., pp.761,-766.

[10] R Core Team, "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, 2013. [Online]. Available: http://www.R-project.org. [Retrieved: Feb. 2015].

 ISBN: 978-1-61208-411-4

Providing Response to Security Incidents in the Cloud Computing with Autonomic Systems and Big Data

Kleber M. M. Vieira, Daniel S. M. Pascal Filho, Carlos B. Westphall, João Bosco M. Sobral, Jorge Werner
{kleber,westphal,bosco}@inf.ufsc.br, {jorge.werner,daniel.smpf}@gmail.com
LRG - INE - UFSC - Florianopolis - SC - Brazil

***Abstract*—This article provides a real-time intrusion response system in order to reduce the consequences of the attacks in the Cloud Computing. Our work proposes an autonomic intrusion response technique that uses a utility function to determine the best response to the attack providing self-healing properties to the environment. To achieve this goal, we propose the Intrusion Response Autonomic System (IRAS), which is an autonomic intrusion response system, using Big Data techniques for data analysis.**

Keywords–Cloud; Security; Intrusion Detection System; Big Data.

I. INTRODUCTION

As a complement to the work presented in [1], the object of this article is to present the results and details of its implementation. Because of their distributed nature, cloud computing environments are a great target for intruders interested in exploring possible vulnerabilities in their services and consequently using the abundant resources maliciously.

The growing number of attacks and vulnerability exploitation techniques requires preventative measures by system administrators. In this context, the need for a highly effective and rapid reactive security system gains importance. These measures are getting more complex with the growth of data heterogeneity and the increasing complexity of the attacks. In addition, slow reaction time from human agents and the huge amount of data and information generated, makes the decision making process an arduous task. In response to this, there is an increase in the usage of Intrusion Detection Systems (IDS) [2], as a way to identify attack patterns, malicious actions and unauthorized access to an environment [3].

The need for IDS is growing due to limitations in Intrusion Preventing Systems (IPS) - which focus on alerting administrators when a vulnerability is detected, connectivity and threat evolution, as well as the financial appeal of cybercrime [4].

Despite their growing importance, currently available IDS solutions have limited response mechanisms. While the research focus is on better intrusion detection techniques, response and effective threat reaction are still mostly manual and rely on human agents to take effect [5].

Recently, some intrusion detection tools have begun providing limited sets of automated responses, but with the growing complexity of intrusions, the need for more effective response system strategies has increased. Due to implementation limitations, research on intrusion detection techniques advance faster than intrusion response systems [3].

The development of reliable and rapid responsive systems is even more important for cloud computing, in which elasticity increases the risk and costs of an attack [6].

A. Motivation

The number of computer attacks has grown in quantity and complexity in the recent years, making defense an increasingly arduous task. Every computer that suffers an attack has very limited information on who initiated the attack and its origin. Current intrusion detection and response systems do not keep up with the growing number of threats [5].

The focus on manual processes creates a delay between detection and response, leaving a window of opportunity for attackers [7]. Research findings by Lumpur [5] indicate that if a skilled attacker has a period of 10 hours between intrusion and response, the attack has an 80% chance of success. If the attacker has 20 hours, the attack has a 95% chance of success, and at 30 hours the attack becomes virtually foolproof. In this situation, the system administrator's skills become irrelevant. On the other hand, if the response to the intrusion is immediate, the chance of a successful attack is almost zero. Lumpur says that statistics have shown that the number of pro-rated intrusions is growing. The high cost of a contract indicates serious financial commitment made by the Pentagon to prevent and secure their infrastructure from another country.

An automated intrusion response system that incorporates the best intrusion detection techniques would offer the best possible defense in a short time frame, affording the system administrator more time to develop a permanent solution to prevent future attacks or to fix the exploited vulnerability [5] [7].

According to Buyya [8], the Cloud is complex, extensive, heterogeneous, and challenging to manage. This environment requires an automated and intelligent system to provide cost-efficient security services. Thus, cloud systems represent a distinct structure, with several layers of abstraction, that requires specific IDS and response techniques to address its complexity.

B. Goals

In this article, we propose a model for autonomic intrusion detection system based on the autonomic loop, commonly referenced as MAPE-K (Monitor, Analyze, Plan, Execute and Knowledge Base). To monitor and analyze, we use sensors to collect data from IDS logs, network traffic, system logs, and data communication. For storage and further analysis, distributed storage is used. For instance, we chose Apache

 ISBN: 978-1-61208-411-4

Hadoop as a storage engine because of its performance, scalability and further capabilities to be extended and perform MapReduce jobs.

This paper is organized as follows: Section 2 describes the proposal's underlying concepts and key technologies. Section 3 presents an overview of the related work. Section 4 details the proposal. Section 5 show the results of execution tests. Section 6 concludes the paper with future goals and open challenges.

II. AUTONOMIC COMPUTING

Autonomic computing can overcome the heterogeneity and complexity of computing systems and is being considered a new and effective approach to implement complex systems, by addressing several areas in which humans are losing control due to system complexity and slow reaction time, such as the security systems area [9].

The autonomic computing model is based on the so called self properties. The self is inspired by the autonomic nervous system of the human body, which can manage multiple key functions through involuntary control. The autonomic computing system is the adjustment of software and hardware resources to manage its operation, driven by changes in the internal and external demands. It has four key features, including self-configuration, self-healing, self-optimization and self-protection.

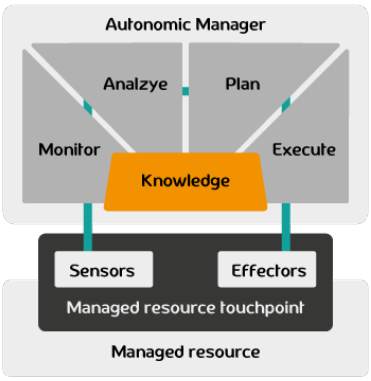


Figure 1. An autonomous system.

Figure 1 shows the structure of an autonomic system and its Monitor, Analysis, Planning, Executor and Knowledge (MAPE-K) cycle [10], composed by the Monitoring, Analysis, Planning and Executing modules. All the management of the autonomic component is performed by a meta-management element, which makes decisions based on the knowledge-base it built.

Sensors are responsible for collecting information from the managed element. Information collected by the sensors is sent to the monitors where they are interpreted, preprocessed, aggregated and presented in a higher level of abstraction. After this, the analysis phase is executed and planning takes place. At this stage, a work plan is created, which consists of a set of actions to be performed by the executor. Only the sensors and executors have direct access to the managed element. Through the autonomic management cycle, there may be a need for decision-making, and thus the presence of the knowledge base is also necessary [11].

A. Autonomic Systems Properties

The essence of autonomic computing is self-management. To implement it, the system must be self-aware as well as environment-aware. Thus, the system must precisely know its current situation and be aware of the operational environment in which it operates. From a practical standpoint, according to Hariri [11], the term autonomic computing has been used to denote systems that have the following properties:

- Self-awareness: the system knows itself, including its components, their state and behavior.
- Context-awareness: the system must be aware of the context of its execution environment and be able to react to changes in its environment.
- Self-configuring: the system must dynamically adjust its resources based on its status and the state of the execution environment.
- Self-optimizing: the system is able to detect performance degradations and functions to perform self-optimization.
- Self-protecting: the system is able to detect and protect its resources from external and internal attackers, maintaining its overall security and integrity.
- Self-healing: the system must have the ability to identify potential problems and to reconfigure itself in order to continue operating normally.

III. RELATED WORK

In this section, five related papers that we considered important to our research were selected. To evaluate these, five topics were chosen to analyze them: focus on IDS, relation to the Cloud scenario, attack response, self-healing method, and algorithm used.

Chai [12] presents an in-flow event processing system for autonomic computing . This system is resistant to hardware failures and attacks. The mechanism votes on consuming events. It also introduces an evidence-based safe-guarding mechanism that prevents a faulty event.

Wu [13] proposes an autonomous manager which introduces a mechanism for multi-attribute auction. Its architecture has a layer of managed resources generically covering all physical devices such as routers, servers and software applications. These resources should be manageable, observable, and adjustable. The state of resources refers to all data (events) that reflect the state of existing resources, including logging and real-time events. This architecture also has an autonomous agent as a detection engine, optimization strategy, autonomic response, and a knowledge base module. Wu says that the autonomic response depends on a knowledge base of possible actions. It is necessary to create a knowledge base with attributes and valuations [13].

The Kholidy [14] approach describes how to extend the current technology and IDS systems. His proposal is based on a hierarchical IDS to experimentally detect DDoS, host-based, network based and masquerade attacks. It provides capabilities for self-resilience preventing illegal security event updates on data storage and avoiding single point of failure across multiple instances of intrusion detection components. Kholidy's proposal consists of a hierarchical structure, autonomic and Cloud based, extending his earlier work with features such as

 ISBN: 978-1-61208-411-4

autonomic response and prediction. In particular, it assesses vulnerabilities and risks in the system through a mechanism that builds a security model based on risk assessment and security event policies criticality. It also provides the possibility of an automatic response to actions based on a set of policies defined by the system administrator. However, a black box format does not clarify possible answers or make clear how to choose the best answer leaving that decision to a system administrator. Finally, the architecture offers some predictive capabilities based on Holt-Winters algorithm [15], which predicts and detects abnormal behavior in network traffic when the amount of collected network traffic is either too high or too low, compared to normal network traffic. Predictive capabilities improve detection accuracy of both decision making and automated response [16].

Vollmer [17] describes new architecture that uses concepts of autonomic computing, based on SOA and an external communication layer to create a network security sensor. This approach simplifies the integration of legacy applications and supports a safe, scalable, self-managed structure. The contribution of this piece is a flexible two level communication layer, based on autonomic computing and SOA. One module uses clustering and fuzzy logic to monitor traffic for abnormal behavior. Another module passively monitors network traffic and deploys deceptive hosts in the virtual network. This work also presents the possibility of an automatic response but it does not address the topic in detail, leaving it for future research.

Sperotto [18] presents an autonomic approach to adjust the parameters of intrusion detection systems based on SSH traffic anomalies. Sperotto proposes a procedure which automatically tunes system parameters, and in doing so, optimizes system performance. Their approach was validated by testing it on a probabilistic-based detection test environment for attack detection, on a system running SSH.

A. About the related works

Related papers representing the state of the art attempt to solve the problem of cyber-attacks by proposing intrusion detection mechanisms and increasing detection techniques. Although many of them show the need for automatic responses, none of them go further in this direction. The works of Wu [13] and Vollmer [17] mention the possibility of attack response. However, neither delves deeper into the issue.

Table I shows a brief comparison of the related works, based on the previously described topics.

IV. PROPOSAL

We propose an intrusion response autonomic system (IRAS) based on MAPE-K. Here we will explain each module of system.

A. Proposed system: IRAS Intrusion Responsive Autonomic System

The approach of IRAS follows the method of an autonomic system for intrusion response. The sensors collect log data from the network IDS and host systems. This information is compiled in a Big Data environment [19], preprocessed and placed in a higher level of abstraction, ready to be sent to the analysis and planning cycles of the autonomic loop.

Based on the MAPE-K autonomic loop, IRAS, as shown in Figure 2, their modules are:

- Monitor: data collection from sensors, and storage on Big Data infrastructure.
- Analysis: preprocessing (filtering, aggregation) and analysis.
- Planning: calculation of utility.
- Executor: based on results of the utility function, effective measures will be taken in the system.
- Knowledge: database, built from the monitored and analyzed data, is fed back into the utility based function, weighting the utilities.

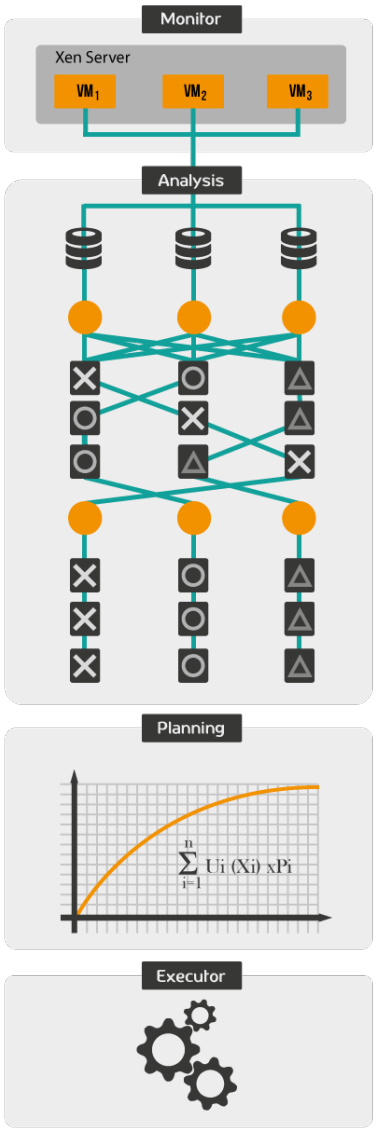


Figure 2. Intrusion Responsive Autonomic System IRAS

B. Monitoring

The first phase of the MAPE-K autonomic cycle corresponds to monitoring. In this step, sensors are used in order

TABLE I. RELATED WORKS

Author	IDS	Cloud	Response	Self-healing	Big Data	Algorithm
Wu	yes	no	yes	no	no	Auction
Kholidy	yes	yes	yes	no	no	Holt- Winters
Vollmer	yes	no	yes	no	no	Fuzzy
Sperotto	yes	no	no	no	no	Flor-based
Chai	yes	no	no	yes	no	Byzantine fault tolerance

to obtain data, reflecting changes in behavior of the managed element, or information from the execution environment that is relevant to the self-management process.

The concept of a sensor is a little generic, but it is possible to consider a sensor as a component of the system that makes the connection between the external world and the management system.

However, the important nuance to observe in data monitoring for security in Cloud Computing is that the data will be intrinsically temporal. This characteristic imposes some peculiarities in the data structure to store temporal information, as well as in the queries to be executed on the sensor database to retrieve useful information.

To monitor and analyze, we used sensors to collect data. It is important collect data from VMs and Hypervisor. Our monitor collects data from IDS logs in the Hypervisor and VMs, network traffic in the entire infrastructure, system logs, and data communication.

C. Analysis

The analysis phase queries the monitoring data looking for events that can characterize attacks.

As defined by Manyika [20], Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. Zikopoulos [21] defines the three data characteristics of Big Data sets: volume, variety and velocity. We have a large volume of data from various sources such as logs, IDS alerts, and network traffic scans, in which processing and analysis speed is necessary to extract meaningful information from these sources. Based on work by Suthaharan [19], we decided to use a structure with Big Data tools, which in this case was Hadoop, to organize the collected data in the Cloud and perform monitoring. However, Suthaharan uses Machine Learning (ML) to find attacks and in this paper we propose to use technical knowledge based on intrusion detection systems [2], making it possible to detect attacks like Stuxnet or Duqu. Thus we made a map reduced over the collected data to identify signatures of known attacks, by extracting significant data such as origin, destination of attack, type, signature and timestamp.

There is a resourceful set of analytics methods that correlates data in order to discover causality relationship, or events association. There are three types of analytical methods that are useful for Cloud Computing security:

- Diagnostic: this method means to synthesize a temporal flow of events arising from sensors in a *security state* of the Cloud - it is common to represent the state as a dashboard.
- Root-cause: the goal of this type of analysis is to determine what events are the main causes of the current Cloud state.
- Prediction: the prediction methods aims to suggest forecast projections to the Cloud state.

It is possible to consider that the analysis phase in Cloud Computing security management has the following characteristics:

- There must be evaluation methods able to supply a set of security metrics for parts of and for the entire Cloud.
- It must consider temporality – generally based on time series.
- It must be multi-criteria – there may be multiple, seemingly uncorrelated, events that, perceived together, constitute an attack.
- It must learn – the measures in a real world Cloud changes their statistical distribution, variance and behavior – in this context, an analytical method to security in Cloud Computing must be adaptive to follow these changes.

The root-cause analysis will not be addressed in this paper. However, it may be important to correlate and determine how some configuration states (e.g. a blocked ip address in the firewall) influence the occurrence of security incidents. In this way, a sensor component reads the data from logs, IDS agents, VM and Hypervisor [22] data collectors, network traffic sniffers, SNMP agents and alarms. This analysis will be important to determine and discover possible security actions.

The prediction will be important to establish the consequences of an action $a \in A$ execution, where A is the set of all possible actions, over a state $s \in S$. So, the prediction of action consequences must provide a probability function $p(s^{t+1}|a, s^t)$, read as: the probability of action a, executed over a state s^t in time t conduce to a state s^{t+1} in time $t+1$.

D. Planning

The Planning Phase receives events from the analysis phase and must choose one action to offer the autonomic system properties self-configuration, self-healing, self-optimization, and self-protection.

To carry out the planning, the Expected Utility technique was chosen [23].

E. Utility

Here we consider the use of utility to find the best response to the attacks. The utility function comes from economy studies (REF) and is expressed by the equation Ui (S). The states that have the best utility should be chosen.

The higher the U, the better. The utility function is expressed as follows:

 ISBN: 978-1-61208-411-4

$$U[x_1, x_2, x_3...x_n] = u_1(x_1)+u_2(x_2)+...u_n(x_n) = \sum_{i=1}^{n} u_i(x_i)$$

$$\max_{x \in D} u[x_1, x_2, x_3...x_n]$$

An example of the application of utility: Let us say that in a meal the utility of coffee is 1, orange juice, 2, bread, 3 and a cookie, 4. Thus, we can express the utility of breakfast by: U (drink, solid) = u. The option with the highest utility should be chosen, which in this case would be U (orange, cookie) = 6.

F. Expected Utility

Incrementing our utility function with the uncertainty that the response may block an attack and bring self-healing to the environment, we use the probability of the event [23].

$$UE[x_1...x_n] = u_1(x_1) \times p_1 + ...u_n(x_n) \times p_n = \sum_{i=1}^{n} u_i(x_i) \times p_i$$

$$\max_{x \in D} u[x_1, x_2, x_3...x_n]$$

For example, given a scan attack, one possible response is to block the source IP.

The probability of this event succeeding is 50%.

$$P(blockIP) = \frac{2}{1}.$$

If the value of the block IP action has a utility value of 5, we can express this as follows:

$$UE(blockIP) = 5 \times 0,5 = 2,5.$$

With the history of this response it is possible to over time optimize the environment, granting the self-healing autonomic property to the environment.

G. Executor

After calculating the response with the highest expected utility, it is possible to forward the response to an executing agent in the Cloud. The hypervisor is responsible for executing the response, providing transparence for each virtual machine.

As shown in Table 1, our work presents an increase in the state of the art when you use Big Data to locate attack occurrences and to be able to provide a response that takes into consideration the impacts of the attack across the Cloud environment. Regarding the authors Wu [13] and Vollmer [17], the contribution of our research was to consider the Cloud environment and the peculiarities of its hypervisor, and the complexity of providing a response without being invasive to customers. Our work also considers self-healing and uses a statistical function in expected utility to achieve the most efficient response and thereby, block the attacks.

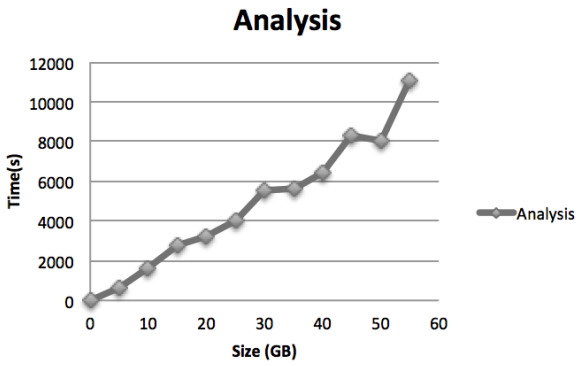


Figure 3. Analysis module execution

V. Execution

We developed a prototype to evaluate the architecture described in this work. It uses Cloudera, Xen Cloud and Cloud Stack. We use JnetPCap to capture network traffic and the parse data. Afterwards we used MapReduce to organize the data by source IP, transport layer and application layer. With organized data the search can be completed more efficiently. After this process is done, a search for known attacks is performed. Data tables are created to perform the experiments with audit elements coming from both the system log and data captured from network. We prepared two types of simulation data to perform the tests.

- Data representing legitimate actions: A set of known services simulating normal behavior was executed to prepare this type of data.
- Data representing knowledge attacks.

The Analysis and Planning module presented in this paper was implemented in Java. For the Analysis module, we used Hadoop. This module was the critical processing point. To perform the MapReduce, 1841 seconds were needed to process 10 GB. The results are shown in Figure 3. After the MapReduce, the result was a small table with data for the Planning module. In this test we used only one instance of the Cloud. To achieve improved performance we could create a larger number of instances.

VI. Conclusion

This paper proposed an autonomic computation system to respond to attacks. The current state of the art was researched and compared with the proposal described here. The solution was distributed into four main modules: Monitoring, Analysis, Planning and Execution. A prototype was presented, which, for the Monitoring module, captured all data transferred in the network. For analysis, we used the Big Data Hadoop tool. For the Planning module, in order to make the best attack response decisions the expected utility function was used, a technique inspired by economics. This solution makes it possible for the Cloud environment to have a self-healing capability against attacks. Tests were performed in the Cloud environment with a large volume of data. The results made it possible to detect attacks and provide a response to them. In this way a we created a self-healing property for the cloud environment. For future research, we suggest focusing on the need to improve the

 ISBN: 978-1-61208-411-4

performance of the Analysis module in order to have a greater efficiency of resource use, in relation to the large amount of data. It is also possible to use a resource limit criterion for the utility function, to get the best response, which uses fewer cloud computing resources.

REFERENCES

[1] K. M. Vieira, F. Schubert, G. A. Geronimo, R. de Souza Mendes, and C. B. Westphall, "Autonomic intrusion detection system in cloud computing with big data," in The 2014 International Conference on Security and Management (SAM 2014)., 2014.

[2] K. Vieira, A. Schulter, C. Westphall, and C. M. Westphall, "Intrusion detection for grid and cloud computing," It Professional, vol. 12, no. 4, 2010, pp. 38–43.

[3] N. Stakhanova, S. Basu, and J. Wong, "A taxonomy of intrusion response systems," International Journal of Information and Computer Security, vol. 1, no. 1, 2007, pp. 169–184.

[4] C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, and M. Rajarajan, "A survey of intrusion detection techniques in Cloud," Journal of Network and Computer Applications, vol. 36, no. 1, Jan. 2013, pp. 42–57. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1084804512001178

[5] K. Lumpur, "An investigation and survey of response options for Intrusion Response Systems (IRSs)," 2010.

[6] P. Mell and T. Grance, "The nist definition of cloud computing (draft)," NIST special publication, vol. 800, no. 145, 2011, p. 7.

[7] C. A. Carver, "Intrusion response systems: A survey," Department of Computer Science, Texas A&M University, College Station, TX, 2000, pp. 77 843–3112.

[8] R. Buyya, R. Calheiros, and X. Li, "Autonomic Cloud computing: Open challenges and architectural elements," Emerging Applications of ..., Nov. 2012, pp. 3–10. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6407847 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6407847

[9] J. Kephart and D. Chess, "The vision of autonomic computing," Computer, vol. 36, no. 1, Jan. 2003, pp. 41–50. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1160055

[10] M. C. Huebscher and J. A. McCann, "A survey of autonomic computing—degrees, models, and applications," ACM Computing Surveys (CSUR), vol. 40, no. 3, 2008, p. 7.

[11] S. Hariri, B. Khargharia, H. Chen, J. Yang, Y. Zhang, M. Parashar, and H. Liu, "The autonomic computing paradigm," Cluster Computing, vol. 9, no. 1, 2006, pp. 5–17.

[12] H. Chai and W. Zhao, "Byzantine fault tolerant event stream processing for autonomic computing," in Dependable, Autonomic and Secure Computing (DASC), 2014 IEEE 12th International Conference on. IEEE, 2014, pp. 109–114.

[13] Q. Wu, X. Zhang, R. Zheng, and M. Zhang, "An Autonomic Intrusion Detection Model with Multi-Attribute Auction Mechanism," vol. 10, no. 1, 2013, pp. 56–61.

[14] H. A. Kholidy, A. Erradi, S. Abdelwahed, and F. Baiardi, "Ha-cids: A hierarchical and autonomous ids for cloud systems," in Computational Intelligence, Communication Systems and Networks (CICSyN), 2013 Fifth International Conference on. IEEE, 2013, pp. 179–184.

[15] C. Chatfield, "The holt-winters forecasting procedure," Applied Statistics, 1978, pp. 264–279.

[16] H. Kholidy, A. Erradi, S. Abdelwahed, and F. Baiardi, "A hierarchical, autonomous, and forecasting cloud IDS," 2013, pp. 213–220.

[17] D. Vollmer, M. Manic, and O. Linda, "Autonomic Intelligent Cyber Sensor to Support Industrial Control Network Awareness," IEEE Transactions on Industrial Informatics, vol. PP, no. 99, 2013, pp. 1–1. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6547755

[18] A.-b. Idss, S. S. H. Case, A. Sperotto, M. Mandjes, R. Sadre, P.-t. D. Boer, A. Pras, and P.-T. de Boer, "Autonomic Parameter Tuning of Anomaly-Based IDSs: an SSH Case Study," IEEE Transactions on Network and Service Management, vol. 9, no. 2, Jun. 2012, pp. 128–141. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6172597

[19] S. Suthaharan, "Big data classification: Problems and challenges in network intrusion prediction with machine learning," in Big Data Analytics workshop, in conjunction with ACM Sigmetrics, 2013.

[20] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, May 2011.

[21] P. Zikopoulos, C. Eaton et al., Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, 2011.

[22] C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, and M. Rajarajan, "A survey of intrusion detection techniques in cloud," Journal of Network and Computer Applications, vol. 36, no. 1, 2013, pp. 42–57.

[23] R. F. Bordley and S. M. Pollock, "A decision-analytic approach to reliability-based design optimization," Operations research, vol. 57, no. 5, 2009, pp. 1262–1270.

 ISBN: 978-1-61208-411-4

Entity Title Architecture Pilot: Deploying a Clean Slate SDN Based Network at a Telecom Operator

Luiz Cláudio Theodoro,
Pedro Henrique A. D. de Melo,
Flávio de Oliveira Silva,
João Henrique de Souza Pereira,
Pedro Frosi Rosa and Alexandre Cardoso

Federal University of Uberlândia
Uberlândia, MG, Brazil
Email: lclaudio@feelt.ufu.br,
pedrodamaso@mestrado.ufu.br, flavio@ufu.br,
joaohs@ufu.br, pfrosi@ufu.br,
and alexandre@ufu.br

Alex Vaz Mendes,
Murilo Borges Gomes Machado,
Hélvio Pereira de Freitas, Luiz Cláudio Theodoro
and João Henrique de Souza Pereira

Innovation, Research and Development
Algar Telecom
Uberlândia, MG, Brazil
Email: alexvaz@algartelecom.com.br,murilo.bgm@algartelecom.com.br,
helvio@algartelecom.com.br, lclaudio@algartelecom.com.br
and joaohs@algartelecom.com.br

***Abstract*—Due to Internet's remarkable success, its architecture is being challenged to attend new applications requirements such as mobility, security and Quality of Experience (QoE), among others. The requirements that the architecture faces today are far away from the design principles of the protocols in the sixties. Several research initiatives are based on a clean-slate approach, a disruptive view that might result in a completely new network. Our research group proposed the Entity Title Architecture (ETArch), a clean-slate Software Defined Networking (SDN) based approach that currently is able to satisfy applications requirements such as support to multicast traffic, mobility and QoE. This work goes further and presents the deployment of ETArch on a telecommunications service provider network. This work contributes to Future Internet initiatives by presenting a viable approach to deploy new network architectures on top of current providers networks.**

Keywords–Software-Defined Networking;Domain Title Service; Workspace; Telecommunications.

I. INTRODUCTION

Despite the huge success, the Internet architecture is facing a completely new technological context that defies its evolution. In spite of its ubiquity, the Internet has some difficulties to attend new applications requirements, such as mobility, security and Quality of Experience (QoE), among others [1]. The developed protocols contributed for the current Internet success but the requirements that the architecture faces today are far away from their design principles [2] of the sixties.

Several research initiatives [3][4] are on their way in order to provide a solution for the new demands regarding the Internet architecture. One of the approaches to evolve Internet architecture is based on a clean-slate view [5], which suggests drastic changes and might result in a completely new network.

In order to experiment with these new network architectures, several infrastructures are being deployed around the world, such as OFELIA [6] in the Europe, GENI [7] in the United States and FIBRE [8] in Brazil in a joint effort with Europe. These infrastructures enable the deployment and the scaling of the experiments that are necessary to face current Internet scale, however ongoing efforts are using infrastructures that are apart from the current Internet.

Although there are several contributions at a global level, it is not easy to reproduce the research outside a laboratory environment. Experiments involving small equipment sets with few users are an important step for a research validation. However, before going to production, this research must be deployed into real infrastructures and the validation must take into account around millions of users. In this scenario, it becomes a critical issue because the companies do not release their plant and environment in order to be manipulated by researchers with the fear of risking themselves because somehow could affect the services provided for the public.

In previous work, our research group proposed the Entity Title Architecture (ETArch) [9], a clean-slate Software Defined Networking (SDN) based approach which aims at satisfying different applications requirements, such as support to multicast traffic [10] , mobility [11], and QoE [12].

The present work goes further and presents and details the deployment of ETArch at a telecommunication service provider production network. This initiative represents an important path towards the actual deployment of Future Internet initiatives on real networks.

ETArch architecture guarantees the possibility of implementing a clean slate SDN network, in a telecom operator, by using the concept of horizontal addressing based on titles, an unambiguous designation of an entity. In a traditional scenario, involving residential customers connected to a commercial network, an alteration in their access methods is induced in a non-conventional way. All of this happens without great setbacks for the final user and a ETArch based chat application is executed with traffic monitoring among the participants.

After this deployment, there is a firm intention to promote a set of developments which can attest the efficiency and potentiality of this solution collaborating so that the future Internet can answer to a series of current and future demands.

The remainder of this work is organized as follows: Section II presents an overview of related work about deployment of new network architectures on top of current networks. Section III introduces ETArch concepts and presents the operator infrastructure. Section IV describes the technical aspects of

 ISBN: 978-1-61208-411-4

the deployment. Section V presents some results of the current work and finally, Section VI presents some concluding remarks and future work.

II. Related Work

Several research groups has initiatives in order to make Internet capable to support the new requirements that challenges the current Internet architecture. One of the approaches is to decouple the architecture and the infrastructure and the OpenFlow standard is cited as one of the most popular solutions to this end [13][14][15]. Some of these initiatives goes towards the direction to deploy the research results onto real environments.

BeHop [16] has an interesting approach, by implementing a wireless testbed for dense WiFi networks frequently found in residential and corporative environments. This prototype was implemented in a campus with about thirty active users as "guinea pigs" allowing researchers to study and evaluate pros and cons of new ways of controlling WiFi networks. BeHop supplies a general purpose framework for experimenting new techniques in order to control power, channel allocations and associations.

The integration between BeHop and the production network showed the benefits of an implementation in the actual world while maintaining the aimed flexibility to process others experiments. It was essential to study and to evaluate the WiFi management strategies and its impacts on the conditions found in a real network such as clients diversity, mobility, and interference with neighbor networks. A testbed was used in order to transport real traffic of users connected to WiFi devices and at the same time to keep the flexibility to apply frequent changes and occasionally force the network down aiming to show network resilience.

Yiakoumis' work [17] points out the possibility of leaving the network control to the user instead of the Internet Service Provider (ISP). This statement raised controversies and contrary opinions have been coming up but the defended idea is to allow that the user's choice can guide the network traffic managing not only inside residences but also inside the ISP [18].

An interesting implementation has been made by Hampel, Steiner and Bu [19]. They suggest the idea of the SDN in an operator, but on top of an Internet Protocol (IP) network. In this case, OpenFlow capable elements run vertical forwarding to interoperate with a legacy infrastructure using IP in consonance with routing traditional protocols.

All these proposals have in common the approach that new solutions could be created by using current network infrastructure. These new solutions are decoupled from the network infrastructure and enables new types of experiments using a SDN based approach.

This study also is based on the assumption that SDN is the enabler of changes in the network which would make it more programmable and flexible. However, it goes further and deploys a clean-slate network architecture on top of a legacy infrastructure at a real network operator. A particular feature of ETArch is that it aims at supporting the application communication requirements over time and support these requirements from top to lower layers of the protocol stack.

III. Entity Title Architecture (ETArch) Pilot

Countless researches are being made with the intention of recreate the Internet architecture that collaborates for the evolution of this great worldwide network. The more expressive proposals have built a large-scale experimentation facility, supporting both research on networks and services, by gradually federating existing and new testbeds for emerging or future Internet technologies [20][4]. Joining this researchers initiative around the world, the ETArch Pilot group intends to create conditions so that researches in future experiments leave the laboratory and go to the actual world.

The suggested architecture in this work has as basic point of view to semantically approximate upper and lower layers and for that it uses the ETArch proposal operating over a commercial network. When it comes to this model it foresees the possibility of attending Internet demands whether current or future ones. ETArch is a clean slate network architecture in which identification and addressing schemes are based in an independent topology designation that uniquely identifies an Entity: its Title. ETArch transport mechanism is based on a logical channels named Workspace which is capable of unifying multiples communication entities.

Entities in the Title Model [21] differ from the defined concept in some literatures and they are not considered simple resources inside a network but beings whose communication needs must be understood and supported by the Service Layer and then by the lower layers as Physical and Link layers. Hereinafter ETArch main concepts and components are better detailed.

An Entity has list of requirements and capacities related to communication. An entity may be an equipment, a user, an application, a thing, and so on. It has at least one title and one location, known as Point of Attachment (PoA). From mobility point of view such separation is important because an entity's location could change over time. These entities can relate among themselves and through such relations they can inherit properties, except the title [9][22].

A group of Titles is bundled in a Namespace which also must have a single title. A Title can be represented by a tuple and its specification could be such as Namespace::Identification-entity. On the other hand, Workspace is a logic bus, independently of underlying topology, with which entities can be attached to be part of a communication domain. The entities addressing happens in an application level and these entities do not communicate directly but they communicate through a workspace.

The Workspace can work on wired and wireless networks. The workspace has the following properties: a title; a group of Network Elements (NE); a list of capacities; the visibility; and a list of requirements. The title identifies that workspace in a unique way. The NE list represents the physical infrastructure that supports the workspace. The capacities indicate the properties that the workspace must satisfy regarding communication such as QoS and security parameters. The requirements must be supported by one entity that wants to attach to a given workspace.

A Workspace is created when an entity produces some kind of thing that can be consumed by other entities such as in a file sharing, a content or a video streaming. It is controlled by a Domain Title Service (DTS), which has the responsibilities

 ISBN: 978-1-61208-411-4

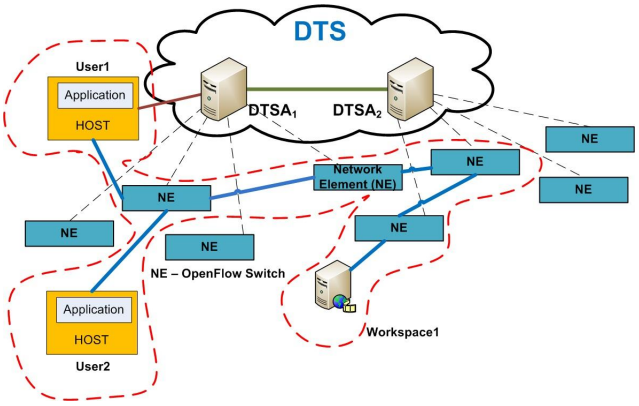


Figure 1. Components of the Architecture - DTS, DTSA, Entity, Title and Workspace.

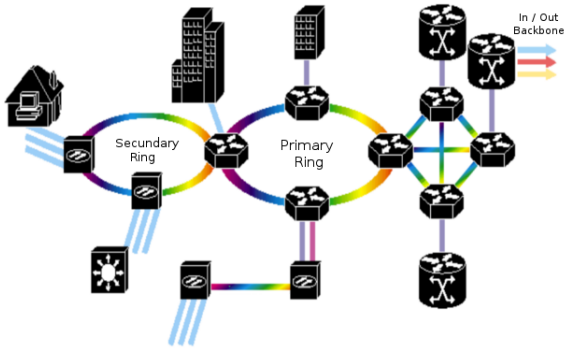


Figure 2. MetroEthernet Topology

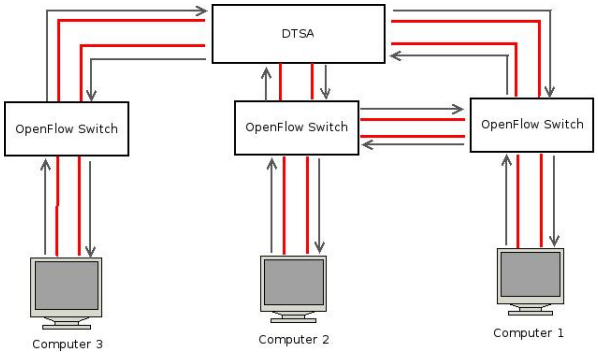


Figure 3. OpenFlow Network

mentioned as follows: titles resolution, entities management and their relationship with their Entities and Titles relation management. The DTS is constituted by DTS Agents (DTSAs) and it has a base of knowledge with information from the environment so it is possible to monitoring and controlling the requisites from an entity. The DTSA manages the entities cycle of life from the beginning to the activities end. The Figure 1 shows an illustration of a Workspace controlled by DTSAs encompassing some NEs.

IV. Deployment and Experimentation

This work proposes the deployment of a testbed based on ETArch architecture as a response to the challenges regarding future internet. By considering that current Internet is an ubiquitous architecture, it is desirable that any changes would be transparent to the users around the world.

By having that ETArch can be deployed on openflow switches, it is possible to implement it by using a telecom operator MetroEthernet network. At the last mile, OpenFlow based switches will be deployed as customer premise equipments.

The operator where this work was conducted is Algar Telecom [23]. A Brazilian operator located in the southeast region of the country. Considering 2014 information, the company has 1.321 million customers and 380 thousand customers in the broadband access. The deployment in this work considers the technologies related to this group of broadband access users.

The MetroEthernet network topology is based on primary and secondary rings as depicted in the Figure 2. Home users are connected to the secondary rings. The primary rings are distributed over the interconnected area and link the secondary rings. Such equipments in their vast majority operate with Synchronous Digital Hierarchy (SDH) or Dense Wavelength Division Multiplexing (DWDM) [24]. Primary rings are characterized by throughput superior to 100 Gbps. Secondary rings link smaller geographic area and their throughput are under 40 Gbps.

To provide access to the customers in the last mile, the operator uses different access technologies such as Asymmetric Digital Subscriber Line (ADSL), Very-high-bit-rate Digital Subscriber Line (VDSL) and Fiber To The Home (FTTH) [25]. Usually the customers are connected to the secondary rings. In some cases, the customers requires higher throughput rates such as in a private condominium, a commercial building or a big company. These customers can be connected to the primary ring as depicted in the Figure 2. In typical ADSL access, the modem is connected to a Digital Subscriber Access Method (DSLAM), which in turn is attached to the secondary ring in a Metro Network. Secondary rings are aggregated in primary ring that connects to the IP routing core.

Two or more areas (domains) will be linked by a virtual networking, by using IP networks, providing researches and developers with an actual usage scenario for the deployments and tests. At a local environment, inside an area, OpenFlow switches are directly linked to each other, as represented by Figure 2.

However, as shown in the operator's network the OpenFlow switches would be separated by an IP structure that does not implement such specification conforming Figure 3.

There are several challenges related with the deployment of ETArch on a MetroEthernet network infrastructure. For example, ETArch does not use the TCP/IP protocol stack on its control and data planes. The legacy infrastructure is completely based on TCP/IP. To solve the problem it was necessary to use a strategy where the ETArch components could communicate in a transparent way over the legacy infrastructure. By using this approach, the legacy infrastructure would work only as a forwarding plane. As a result, the OpenFlow switch will work as being directly connected, as depicted in the Figure 3.

The forwarding graph in an OpenFlow based network can be discovered by using the Link Layer Discovery Protocol (LLDP) [26]. Briefly, LLDP is a link layer protocol used

 ISBN: 978-1-61208-411-4

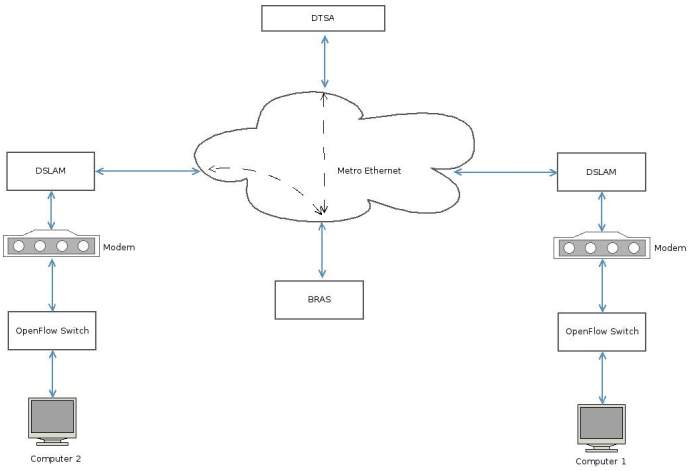


Figure 4. Operator Network

by network devices for informing their identity on a local area network (LAN). The forwarding graph contains all of the OpenFlow switches that are directly connected. However, the legacy infrastructure contains several middleboxes between the switches. These switches are also geographically scattered along the network infrastructure.

Therefore, it was necessary to solve this issue to have all switches directly connected in a virtual way. The adopted solution was to create a tunnel using the Generic Routing Encapsulation (GRE) protocol [27] in order to forward the LLDP frames. The GRE protocol is very popular due to its simplicity and compatibility. GRE is a tunneling protocol that can encapsulate, among other things, the link layer protocols inside a virtual point-to-point link over an IP network. Thus, it was possible to create a virtual network of OpenFlow switches directly connected over the operator IP network, according to Figure 4.

Thus, the tunnel creation could interconnect the OpenFlow switches. However, the telecom operator does not have OpenFlow capable switches deployed on the infrastructure. Replacing current switches is not a feasible alternative. To overcome this, Open vSwitch (OVS) was used. Although this brings a new requirement for the deployment, OVS can be easily installed in commodity hardware at customer premises. OVS is an open-source implementation of a virtual multilayer switch and support multiple protocols and standards including some OpenFlow versions.

In the end, the virtual switches, based on OVS, where scattered along the network and interconnected using GRE tunnels. The DTSA acts as the OpenFlow controller of these switches.

The deployment was based on the most recent version of the software components of ETArch. The components were installed in servers located inside the Operator network. In the first stages, PCs using Ubuntu Linux (14.04) operating systems played the role of each customer's equipment. On each PC, OVS was installed and a new bridge interface (br0) was created. This bridge interface can be used also for data plane and for the control plane in order to communicate with the DTSA.

From the moment this process is realized in both machines, the bridge is added to the switch and established a register in the controller. Both switches (machines) use a GRE because it offers a tunnel, by simulating a link between two network nodes, and this is done by using an IP address (then attending to the initial proposal). Atop of it, the OVS itself is already capable of creating this kind of tunnel internally thus the process is finished by adding the GRE tunnel in the same bridge of the controller.

V. RESULTS

To meet the initially proposed objectives, the verification may be done by the assembly of the network structure and by the execution of an application specifically developed for the ETArch architecture.

The first scan mode is based on the application of ETArch controller which asks every registered switch to send LLDP as a regular package. This proceeding is very important for the network mapping itself because it signals what are the most important options and the paths between them.

In this way, the controller can take a complete view of the network. A packet monitoring tool (tcpdump) has been used to observe the exchange of information between switches, specifically the LLDP messages, encapsulated in GRE packets as mentioned before, as shown in Figure 5.

Such information indicates that these options are virtually adjacent; which means that the legacy structure remains present in the network lab, transparently, where it was initially deployed, which did not require considerable changes in the usual practices within the carrier.

Upon confirmation of adequate controller setup, the switches (machines) were transferred to MetroEthernet environment, where it obtained a valid IP. Every customer has been connected by using the existing network, meaning that customers would be connected to each other by the controller and network elements such as switches, DSLAMs and the access modems. For the construction of GRE tunnels, the OVSs must have a valid IP and therefore the customer must send the Ethernet over the Point-to-Point Protocol (PPPoE) frames to the ISP.

The OVSs setup process is repeated by replacing the previous IPs, by the one obtained on the new network, and thus all the legacy structures are transparent for the control and connection between customers. LLDP packets and their respective answers are inspected to check whether this phase of the process has been successful.

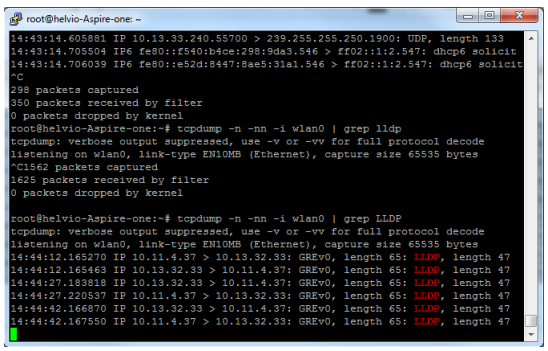


Figure 5. Monitoring LLDP

 ISBN: 978-1-61208-411-4

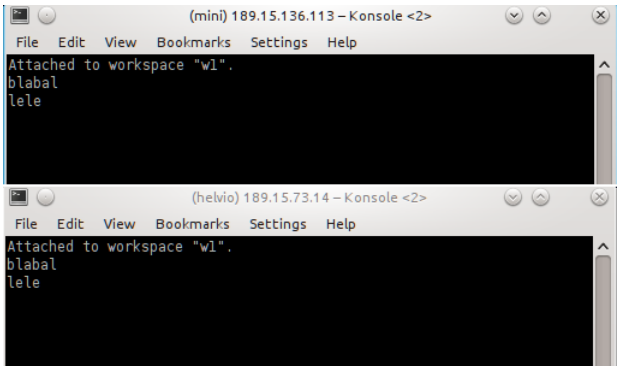


Figure 6. Running Chat

A chat application is invoked to prove that the deployed environment, which involves entities and other ETArch communications concepts[28], is working properly (even though it is using non-traditional TCP/IP protocols).

The chat application has been developed in Python and invoked directly on a command line interface, by passing parameters as entity titles and Workspace. By invoking them (chats) on both machines, each one by its title, it was found that, according to the concept defined in the ETArch architecture, both applications could share the specified workspace (W1), as it can be seen in Figure 6. It allows validate the implementation and shows the results obtained by the new architecture in the proposed environment.

VI. Concluding Remarks and Future Work

This work presented the deployment of ETArch, a clean slate network architecture, over the production network of a telecom operator. In order to test and experiment this deployment, a chat application was used. This application is based on ETArch concepts such as Workspaces, Entities and Titles enabling the use of a new protocol stack between end users over the legacy networks.

In this process, the physical infrastructure was kept in place and only software based framework where added to the infrastructure. In the end user side, a software based OpenFlow switch was introduced and on the operator side, the DTSA, the entity responsible for the control plane of ETArch.

As a future work there are several previewed fronts such as increasing the number of customers and switches, the deployment of new applications based on ETArch workspaces in order to show its efficiency in areas such as video streaming and finally the withdraw of some network elements from the legacy network.

We are confident that this deployment will bring facilities and dynamism to researchers facing the Future Internet's evolution and it can enable new types of services and applications which can be offered by the operator to their customers, helping to bring Future Internet research into reality.

Acknowledgment

This work has been partially funded by the European Community's Seventh Framework Programme, under grant agreement n. 258365 (OFELIA project), by the Brazilian agencies: CAPES, CNPq and FAPEMIG and also by PROPP/UFU. We also would like to thanks ALGAR Telecom for the support and partnership on this work.

References

[1] T. Zahariadis et al., "Towards a future internet architecture," in The Future Internet. Future Internet Assembly 2011: Achievements and Technological Promises, ser. Lecture Notes in Computer Science, J. Domingue, A. Galis, A. Gavras, T. Zahariadis, and D. Lambert, Eds. Berlin, Heidelberg: Springer-Verlag, May 2011, vol. 6656, pp. 7–18. [Online]. Available: http://www.springerlink.com/content/978-3-642-20897-3#section=881237&page=15&locus=86

[2] D. Clark, "The design philosophy of the darpa internet protocols," SIGCOMM Comput. Commun. Rev., vol. 18, no. 4, Aug. 1988, pp. 106–114. [Online]. Available: http://doi.acm.org/10.1145/52325.52336

[3] J. Pan, S. Paul, and R. Jain, "A survey of the research on future internet architectures," Communications Magazine, IEEE, vol. 49, no. 7, July 2011, pp. 26–36.

[4] A. Gavras, A. Karila, S. Fdida, M. May, and M. Potts, "Future internet research and experimentation: the fire initiative," ACM SIGCOMM Computer Communication Review, vol. 37, no. 3, 2007, pp. 89–92.

[5] J. Rexford and C. Dovrolis, "Future internet architecture: clean-slate versus evolutionary research," Communications of the ACM, vol. 53, no. 9, 2010, p. 36–40.

[6] OFELIA. OpenFlow in europe - linking infrastructure and applications. [Online]. Available: http://www.fp7-ofelia.eu/about-ofelia/ [retrieved: May, 2014]

[7] C. Elliott, "GENI: opening up new classes of experiments in global networking [Internet predictions]," IEEE Internet Computing, vol. 14, no. 1, Feb. 2010, pp. 39–42.

[8] FIBRE. FIBRE Project - Future Internet Testbeds Experimentation Between Brazil and Europe. [Online]. Available: http://www.fibre-ict.eu/ [retrieved: May, 2014]

[9] F. de Oliveira Silva, J. H. de Souza Pereira, P. F. Rosa, and S. T. Kofuji, "Enabling future internet architecture research and experimentation by using software defined networking," in Software Defined Networking (EWSDN), 2012 European Workshop on. IEEE, 2012, pp. 73–78.

[10] M. Amaral Gonçalves, F. de Oliveira Silva, d. S. Pereira, J. Henrique, and P. Frosi Rosa, "Multicast Traffic Aggregation through Entity Title Model," Jul. 2014, pp. 175–180. [Online]. Available: http://www.thinkmind.org/index.php?view=article&articleid=aict_2014_7_40_10177

[11] C. Guimaraes et al., "IEEE 802.21-enabled Entity Title Architecture for Handover Optimization," in 2014 IEEE Wireless and Communications and Networking Conference. Piscataway, NJ:: IEEE, 2014, pp. 2671–2676.

[12] Castillo et al., "Evolving Future Internet Clean-Slate Entity Title Architecture with Quality-Oriented Control Plane Extensions," Jul. 2014, pp. 161–167. [Online]. Available: http://www.thinkmind.org/index.php?view=article&articleid=aict_2014_7_20_10164

[13] B. Raghavan et al., "Software-defined Internet architecture: Decoupling architecture from infrastructure," Journal of Something, 2012, p. 43–48.

[14] T. Anderson et al., "NEBULA - A Future Internet That Supports Trustworthy Cloud Computing," 2012, pp. 1–31.

[15] J. Roberts, "The clean-slate approach to future internet design: a survey of research initiatives," annals of telecommunications - annales des télécommunications, vol. 64, no. 5-6, May 2009, pp. 271–276. [Online]. Available: http://dx.doi.org/10.1007/s12243-009-0109-y

[16] Y. Yiakoumis et al., "BeHop: a testbed for dense wifi networks," in Proceedings of the 9th ACM international workshop on Wireless network testbeds, experimental evaluation and characterization. ACM, 2014, pp. 1–8.

[17] ——, "Putting home users in charge of their network," in Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, 2012, pp. 1114–1119.

[18] M. Chetty, R. Banks, A. Brush, J. Donner, and R. Grinter, "You're capped: understanding the effects of bandwidth caps on broadband use in the home," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2012, pp. 3021–3030.

 ISBN: 978-1-61208-411-4

[19] G. Hampel, M. Steiner, and T. Bu, "Applying software-defined networking to the telecom domain," in Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on. IEEE, 2013, pp. 133–138.

[20] J. Rexford and C. Dovrolis, "Future internet architecture: Clean-slate versus evolutionary research," Commun. ACM, vol. 53, no. 9, Sep. 2010, pp. 36–40. [Online]. Available: http://doi.acm.org/10.1145/1810891.1810906

[21] J. de Souza Pereira, F. de Oliveira Silva, E. Filho, S. Kofuji, and P. Rosa, "Title model ontology for future internet networks," in The Future Internet, ser. Lecture Notes in Computer Science, J. Domingue et al., Eds. Springer Berlin Heidelberg, 2011, vol. 6656, pp. 103–114. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-20898-0_8

[22] F. d. O. Silva, J. H. d. S. Pereira, S. T. Kofuji, and P. F. Rosa, "Domain title service for future internet networks," in Anais do II Workshop de Pesquisa Experimental na Internet do Futuro (WPEIF). Campo Grande: SBC, 2011, pp. 33–36.

[23] TELECO. Algar telecom. [Online]. Available: http://www.teleco.com.br/en/en_operadoras/ctbc.asp [retrieved: May, 2015]

[24] M. Huynh and P. Mohapatra, "Metropolitan ethernet network: A move from lan to man," Computer Networks, vol. 51, no. 17, 2007, pp. 4867–4894.

[25] A. S. Tanenbaum, "Computer networks 4th edition," ed: Prentice Hall, 2003.

[26] P. Congdon, "Link layer discovery protocol and mib," V1. 0 May 20. 2002, http://www. IEEE802, 2002.

[27] S. Hanks, D. Meyer, D. Farinacci, T. Li, and P. Traina, "RFC 2784 - generic routing encapsulation (GRE)," 2000.

[28] F. de Oliveira Silva et al., "Semantically enriched services to understand the need of entities," in The Future Internet, ser. Lecture Notes in Computer Science, F. Álvarez et al., Eds. Springer Berlin Heidelberg, 2012, vol. 7281, pp. 142–153. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-30241-1_13

 ISBN: 978-1-61208-411-4