AICT 2013

The Ninth Advanced International Conference on Telecommunications

ISBN: 978-1-61208-279-0

June 23 - 28, 2013

Rome, Italy

AICT 2013 Editors

Michael D. Logothetis, University of Patras, Greece

Mariusz Glabowski, Poznan University of Technology, Poland

Dragana Krstic, University of Nis, Serbia

AICT 2013

Forward

The Ninth Advanced International Conference on Telecommunications (AICT 2013) held on June 23 - 28, 2013 - Rome, Italy, covered a variety of challenging telecommunication topics ranging from background fields like signals, traffic, coding, communication basics up to large communication systems and networks, fixed, mobile and integrated, etc. Applications, services, system and network management issues also received significant attention.

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

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

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

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

We take this opportunity to thank all the members of the AICT 2013 Technical Program Committee as well as the numerous reviewers. The creation of such a broad and high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to

contribute to the AICT 2013. We truly believe that, thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the AICT 2013 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that AICT 2013 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in telecommunications.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the historic charm Rome, Italy.

AICT 2013

Committee

AICT Advisory Committee

Tulin Atmaca, Telecom SudParis, France
Eugen Borcoci, University Politehncia Bucharest, Romania
Michael D. Logothetis, University of Patras, Greece
Go Hasegawa, Osaka University, Japan
Reijo Savola, VTT Technical Research Centre of Finland - Oulu, Finland
Michael Massoth, University of Applied Sciences - Darmstadt, Germany

AICT Special Area Chairs

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

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

AICT 2013 Technical Program Committee

Fatma Abdelkefi, High School of Communications of Tunis - SUPCOM, Tunisia
Sachin Kumar Agrawal, Samsung Electronics, India
Mahdi Aiash, Middlesex University - London, UK
Anwer Al-Dulaimi, Brunel University - Middlesex, UK
Sabapathy Ananthi, University of Madras, India
Pedro A. Aranda Gutiérrez, University of Paderborn, Germany
Miguel Arjona Ramírez, University of São Paulo, Brazil
Andres Arjona, Nokia Siemens Networks, Japan
Michael Atighetchi, Raytheon BBN Technologies-Cambridge, USA
Tulin Atmaca, TELECOM SudParis, France
Konstantin Avratchenkov, INRIA- Sophia Antipolis, France
Paolo Barsocchi, ISTI/National Research Council - Pisa, Italy
Ilija Basicevic, University of Novi Sad, Serbia
Carlos Becker Westphall, Federal University of Santa Catarina, Brazil
Daniel Benevides da Costa, Federal University of Ceará (UFC), Brazil
Ilham Benyahia, Université du Québec en Outaouais, Canada
Robert Bestak, Czech Technical University in Prague, Czech Republic
Eugen Borcoci, University "Politehnica" of Bucharest (UPB), Romania
Christos Bouras, University of Patras, Greece
Lubomir Brancik, Brno University of Technology, Czech Republic
Peter Brida, University of Zilina, Slovakia
Julien Broisin, Université Paul Sabatier, Toulouse III, France

Cristina Oprea, Politehnica University of Bucharest, Romania
Harald Øverby, Norwegian University of Science and Technology - Gløshaugen, Norway
Ali Ozen, Nuh Naci Yazgan University, Turkey
Constantin Paleologu, University Politehnica of Bucharest, Romania
Jari Palomäki, Tampere University of Technology - Pori, Finland
Andreas Papazois, RACTI & CEID / University of Patras, Greece
Cathryn Peoples, University of Ulster, UK
Fernando Pereñíguez García, University of Murcia, Spain
Jordi Pérez Romero, Universitat Politecnica de Catalunya (UPC) - Barcelona, Spain
Maciej Piechowiak, Kazimierz Wielki University - Bydgoszcz, Poland
Michael Piotrowski, University of Zurich, Switzerland
Andreas Pitsillides, University of Cyprus-Nicosia, Cyprus
Adrian Popescu, Blekinge Institute of Technology - Karlskrona, Sweden
Neeli R. Prasad, Aalborg University, Denmark
Emanuel Puschita, Technical University of Cluj-Napoca, Romania
Dusan Radovic, TES Electronic Solutions GmbH - Stuttgart, Germany
Adib Rastegarnia, University of Tehran, Iran
Ustijana Rechkoska Shikoska, University for Information Science & Technology "St. Paul the Apostle" - Ohrid, Republic of Macedonia
Eric Renault, Telecom SudParis, France
Lorayne Robertson, University of Ontario Institute of Technology, Canada
Pawel Rózycki, University of IT and Management, Poland
Danguole Rutkauskiene, Kaunas University of Technology, Lithuania
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
Sergei Semenov, Renesas Mobile Corporation, Finland
Sandra Sendra Compte, University Polytechnic of Valencia, Spain
Dimitrios Serpanos, University of Patras, Greece
Michelle Sibilla, Paul Sabatier University Toulouse 3, France
Nicolas Sklavos, Technological Educational Institute of Patras, Hellas
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
Daniele Tafani, Dublin City University, Ireland
Yutaka Takahashi, Kyoto University, Japan
Yoshiaki Taniguchi, Osaka University, Japan
Richard Trefler, University of Waterloo, Canada
Thrasyvoulos Tsiatsos, Aristotle University of Thessaloniki, Greece
Kazuya Tsukamoto, Kyushu Institute of Technology-Fukuoka, Japan

Kenneth Turner, The University of Stirling, Scotland
Masahiro Umehira, Ibaraki University, Japan
Guillaume Valadon, French Network and Information Security Agency, France
John Vardakas, University of Patras, Greece
Manos Varvarigos, University of Patras, Greece
Dimitris Vasiliadis, University of Peloponnese Greece
Calin Vladeanu, University Politehnica of Bucharest, Romania
Luca Vollero, Università Campus Bio-Medico di Roma, Italy
Krzysztof Walkowiak, Wroclaw University of Technology, Poland
Mea Wang, University of Calgary, Canada
Amali Weerasinghe, University of Canterbury, New Zealand
Steve Wheeler, University of Plymouth, UK
Bernd E. Wolfinger, University of Hamburg, Germany
Mudasser F. Wyne, National University - San Diego, USA
Kang Xi, Polytechnic Institute of New York University, USA
Qin Xin, Université Catholique de Louvain - Louvain-la-Neuve, Belgium
Miki Yamamoto, Kansai University, Japan
Qing Yang, Ciena Corporation, USA
Vladimir S. Zaborovsky, Technical University - Saint-Petersburg, Russia
Giannis Zaoudis, University of Patras, Greece
Smékal Zdenek, Brno University of Technology, Czech Republic
Demóstenes Zegarra Rodríguez, University of São Paulo, Brazil
Liaoyuan Zeng, University of Electronic Science and Technology of China, China
Rong Zhao, Detecon International GmbH - Bonn, Germany
Zuqing Zhu, University of Science and Technology of China, China
Martin Zimmermann, Hochschule Offenburg - Gengenbach, Germany
Sladjana Zoric, Deutsche Telekom AG, Bonn, Germany
Piotr Zwierzykowski , Poznan University of Technology, Poland

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

New Approach to Call Admission Control based-on Interference for Hot-spot Cell

Woogoo Park
Future Technology Research Department
ETRI
Daejeon, KOREA
wgpark@etri.re.kr

Myungae Chung
Future Technology Research Department
ETRI
Daejeon, KOREA
machung@etri.re.kr

***Abstract*—In this paper, we propose a new call admission control (CAC) scheme for wireless cellular systems supporting a hot-spot cell. Two strategies for CAC are assumed. One is based on the number of calls, which are accepted or not, the other is based on the interference level. The proposed CAC can regulate the call attempts with packet by adopting measured interference used for matrix. A function, specifically a linear interpolation function based on Inverse Transform theory is used to convert controlled predictive average admission ratio in a call rejection ratio. Then, a call admission ratio is obtained from the call rejection ratio. Numerical results show that the proposed scheme yields better performance than the conventional scheme does under the heavy calls on a hot-spot cell.**

Keywords-Call Admission Control; Hot-spot Cell; Traffic Model; Interference; Inverse Transform;

I. INTRODUCTION

In the future wireless communications, the main problem is focused on how to efficiently utilize the limited frequency spectrum. The CAC schemes have been mainly developed in switching system maintained by stored program control. The system is formed with a plurality of processors, which are distributed and are formed in a hierarchical structure. The base station performs a predetermined call control process and a non-call control process in accordance with the internal process. In addition, a designated protocol is performed through an interaction between base stations. When interference of a base station is increased, such a load affects other base stations and the entire cells. As a result, the service of the system may be interrupted thereby. Hot-spot activation scheme in classical wire telephone is based on weighted interference. The conventional call admission schemes are directly used by the weighted interference so as to ensure the rapid response of control [1]. However weighted interference provides only rough information on the predictive offered calls, because the offered load is caused by a call. To reduce this degradation, precise CAC scheme based on interference is required. This paper analyzes the effect of weighted interference and call fail ratio on the system performance in theory. Weighted interference efficiency equations are derived for both systems with and without CAC. Analysis results show that call admission can get maximum call efficiency while adjusting call admission can improve call fail for systems without CAC.

This paper is organized as follows: In section II, related works have been studied and in section III cell, traffic and system model are proposed, and also we classifies calls in terms of their characteristics. Section IV describes the call admission activation scheme using interference and call success ratio considering matrix, and also shows the theoretical backgrounds by re-analyzing the relation between interference and wireless channel resources. In section V, we evaluate the proposed CAC model. Finally, Section VI presents the conclusion.

II. RELATED WORKS

The capacity of a cell is limited by its interference and channel resource, which depends on offered calls [1]. The CAC schemes associated with communication systems have been mainly developed on switching system maintained by stored program control [2]. Recently CAC has been dealt with for data service only [3], [4]. CAC is necessary to guarantee the quality of service (QoS) for the users in service, which is mainly concentration on a hot-spot area. In hot-sot area, cell poses significant challenges because of interference, user mobility, and limited channel resource. To efficiently accept call attempts and effectively maximize cell capacity, the cell size is getting smaller. However the smaller cell size is, the higher interference is. So we need to improve the call admission ratio in bad interference environments.

The CAC thresholds derived from the traffic theory [5] and [10]. In [10], the tri-threshold bandwidth reservation scheme supports multi-class services bearing differentiated QoS requirements. It is directed to a mobile switch formed with processes for a judgment such as an admission, a control and a release. We used weighted interference as traffic load for CAC. A QoS-aware wireless MAC protocol called Hybrid Contention-Free Access (H-CFA) and a VoIP call admission control technique called the Traffic Stream Admission Control (TS-AC) algorithm has been presented in [11] to support VoIP, specially on WiFi technologies which have already matured commercially.

It is practically, however, hard to apply to the real-time system. In this paper we suggest that CAC scheme safely maintain the call service by analyzing the characteristics of the weighted interference and the call fail ratio due to congestion, and increases the call

 ISBN: 978-1-61208-279-0

admission ratio [6]. We deal with data traffic as well as voice.

III. PROPOSED MODEL

A. Cell Model

Fig. 1 depicts a cluster of 19 cells being built around interfering base stations. Base station density is based on [9] (cell radius of 4 km for rural/macro and 1.5 km for urban/macro have been assumed). Some comparative simulations were also performed with cell radio as low as 500m and as high as 9 km. The contribution from interferers beyond the closest 19 is considered to be insignificant.

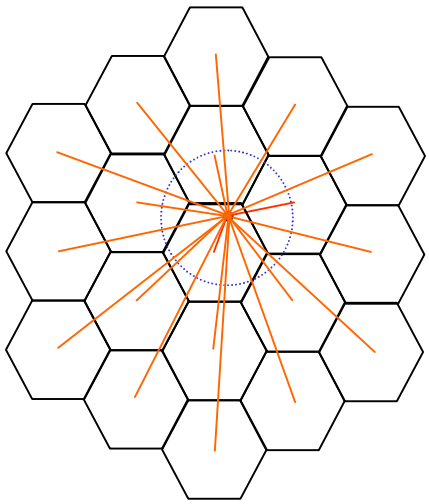


Figure 1. A Cluster of 19 interfering cells

B. Traffic Model

We consider that the call arrival process at the base station is Poisson [7]. Assuming that the call generating process of all the users in a hot-spot cell are independent, the call arrival process from each and every user is also Poisson [8], with a mean call arrival ratio λ , where $\lambda = \lambda^n + \lambda^h$.

TABLE I. CHARACTERISTICS OF CALLS

Location	Service Types	Direction	Arrival Ratio	Priority
Inner Cell	Originating Call	No	λ^n	4
	Terminating Call	No	λ^n	1
Inter Cell	Incoming Handover Call	From Other Cell	λ^h	3
	Outgoing Handover Call	To Other Cell	λ^h	2

The movement of each user is modeled by the two-dimensional random walk and is stationary process including handover from/to neighboring cells. User's channel holding time follows a negative exponential distribution with mean $1/\mu$.

The call admission activation function periodically audits the intensity of call, weighted interference, and call success ratio by time t, in order to avoid the degrading of the performance of system and service ratio. Gradually, a grading limitation of execution module is performed in accordance with the call admission status. The following calls are commonly used to simulate call admission situations [6]:

- Incoming call (new calls)
- Incoming handover call (from other cell)
- Outgoing handover call (to other cell)
- Terminating call

Table I shows the characteristics of call services according to the location-generated services.

C. System Model Description

Considering hot-spot area in base station, some strategies for base station on which calls may concentrate is needed. To obtain the current interference level, a call admission activation scheme must be able to manage interference. For a simple expression of the calculations latter, we use a matrix presentation. The weight sequence in base station can be written as a vector by

$$w_b = [w_1, w_2, w_3, \ldots w_n] \quad (1)$$

where n denotes the number of base stations. From the above, we can see that w_b is a simple function of weight distribution. When weight satisfies normal distribution, weighted interference the cell can be written as

$$\rho_i = I \cdot w \quad (2)$$

Equation (2) constructs an equation between interference and the weight sequence. In order to increase call be accepted, it is assumed that each of interferences of cells is required. Interference among base stations can be written as a matrix by

$$I = \begin{bmatrix} i_{11} & i_{21} & \cdots & i_{n1} \\ i_{12} & i_{22} & \cdots & i_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ i_{1m} & i_{2m} & \cdots & i_{nm} \end{bmatrix} \quad (3)$$

where $i_{11}, i_{21}, \ldots i_{n1}$ are the interference of cells in an base station and m is the number of base stations. Assume that the system has been perfectly synchronized at the base station. The term ρ_i can be denoted by,

$$\begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_m \end{bmatrix} = \begin{bmatrix} i_{11} & i_{21} & \cdots & i_{n1} \\ i_{12} & i_{22} & \cdots & i_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ i_{1m} & i_{2m} & \cdots & i_{nm} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \quad (4)$$

where ρ_i can be simply obtained by

 ISBN: 978-1-61208-279-0

$$\rho_i = \frac{\sum_{k=1}^{m} I_k}{m} \quad (5)$$

D. Relationship between Weighted Interference and Call Success Ratio

To obtain the current weighted interference level described in the (5), a call admission activation scheme must be able to manage information resulted from the following sources primarily:

- call processing
- handover control
- weighted interference

Letting $\rho_{i\_\max}$ and ρ_{i_other} be the acceptable total interference and interference from other cells, respectively, the communication quality is satisfied [5].

$$E_b((c+h)-1)/pg + N_0 + \rho_{i\_other} \le \rho_{i\_\max} \quad (6)$$

where E_b, c, h, pg, and N_0 are bit energy, the number of active calls and handover calls in the same cell, processing gain, and thermal noise power density, respectively. In this paper, E_b, pg, and N_0 are used as reference values. The weighted interference can be adjusted by the use of equation (6). $\rho_{i\_cur}(t)$ is current weighted interference activated by time t, which is given by,

$$\rho_i = \rho_{i\_cur}(t) = \rho_{i\_prev} + \rho_{i\_fail}(\rho_{i\_\max} - \rho_{i\_prev}) \quad (7)$$

where ρ_{i_prev} and ρ_{i_fail} are previous weighted interference and call fail ratio per a base station during inter call admission monitoring time (τ),. We assume a decision scheme where activation time is divided into equal periods- τ . When call admission is increased suddenly, the ratio of call admission is decided by the comparison of the previous call admission-ratio with present one. Using $\rho_{i\_cur}(t)$, the adaptive average weighted interference (AAWI, ρ_{i_aawi}) has very slowly or rapidly increased makes call admission can be expressed,

$$\rho_{i\_aawi}(t) = \frac{\rho_i + \rho_{i\_aawi}(t-\tau)}{2} \quad (8)$$

where $\rho_{i\_aawi}(t-\tau) < \rho_{i\_aawi}(t)$ for $(t-\tau < t)$. The adaptive average weighted interference naturally makes the activation of rapidly increasing call admission possibly. The following (9) shows the parameter of a heavy transient call admission to activate for $(t-2\tau),(t-\tau)$, and t , respectively. We are interested in the degree of gradient-based on offered load as can be found in (9) and (10). From a gradient of rapidly increasing call admission $(D = \frac{\Delta y}{\Delta x})$, the following equation can be

$$\Delta x = \rho_{i\_aawi}(t-\tau) - \rho_{i\_aawi}(t-2\tau) \quad (9)$$

$$\Delta y = \rho_{i\_aawi}(t) - \rho_{i\_aawi}(t-\tau) \quad (10)$$

where $\rho_{i\_aawi}(t-2\tau) < \rho_{i\_aawi}(t-\tau)$ for $(t-2\tau < t-\tau)$. We are interested in the degree of gradient based on offered calls as can be found in (9) and (10). Now, we introduce variable ΔD that indicates whether or not the equivalent number of the degree of rapidly increasing call status D exceeds 2.

$$\Delta D = \begin{cases} 1.2, for\ D > 2 \\ 1.0, for\ D \le 2 \end{cases} \quad (11)$$

It is hard to predict the flow of load because of user mobility and social events, such as huge accidents, bargain sale, and sports game watching. To solve the problem, we introduce adjusted predictive average weighted interference (APAWI) as the optimization parameter (ρ_{i_apawi}). The adjusted predictive average weighted interference is generated by adaptive average weighted interference with ΔD .

$$\rho_{i\_apawi} = \rho_{i\_aawi}(t) * \Delta D \quad (12)$$

The call admission status will be detected when $\rho_{i\_apawi} > 50$.

IV. CAC MODEL

A. Call Admission Activation Scheme

The call admission activation scheme is an autonomous call admission-monitoring module to calculate interference on a base station. As a call admission activation scheme starts, it measures weighted interference according to the cell situations. It also shows the scheme showing the entire detecting of the process of CAC feature, which is formed by an initialization, a judgment, activation, and a release. As heavy call attempts disappear, call admission state is released and the system returns to the normal state and the following four cases will happen;

1). Initialization of call admission activation: upon starting this call admission activation scheme, stores the c urrent weighted interference for adaptive average weighte d interference after call admission monitoring time interv al,

2). Reanalysis of weighted interference: the scheme selects the blocking ratio of call services and calculates a new value of weighted interference with maximum interf erence. After calculating a new weighted interference, the current adaptive average weighted interference sets up b oth new weighted interference and previous adaptive aver

age weighted interference,

3). Adjustment of heavy transient status: when the a daptive average weighted interference from the weighted interference including blocking ratio of call services can be calculated, the gradient is adjusted to adopt a heavy tr ansient status,

4). Activation of call admission situations: finally, we get the adjusted predictive average weighted interfere nce by multiplying adaptive average weighted interferenc e by the degree of heavy transient status and we activate t he call admission according to the adjusted predictive av erage weighted interference.

B. Inverse Transform for Call Admission Ratio

The call admission ratio is generated that adjusted predictive average weighted interference compares with interval for basic call admission ratio (BCAR) and linear interpolation is used for calculating basic call admission ratio. The basic call admission ratio has a capability for admitting a call according to weighted interference. The real underlying distribution, $F(x)$, of basic call admission ratio (the lower line curve in Fig. 2) can be estimated by the empirical cdf, $F'(x)$ (the upper line curve in Fig. 2). The particular curve in Fig. 2 illustrates one changed shape of this underlying distribution, and also that $F'(x)$ is an estimate of $F(x)$.

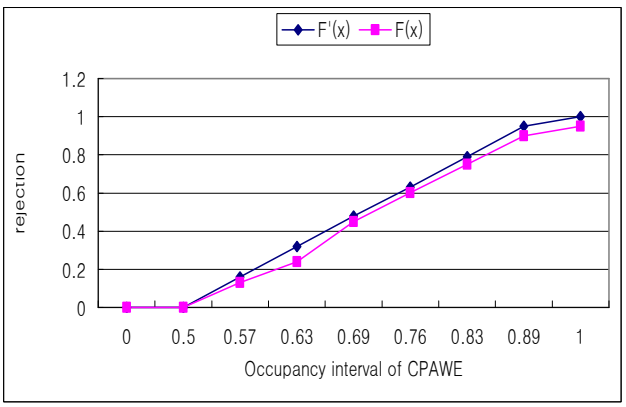


Figure 2 A diagram of Empirical ($F'(x)$) and theoretical ($F(x)$) distribution functions for occupancy interval of adjusted predictive average weighted interference

The empirical cdf $F'(x)$ is defined by using the information in Table II. Each interval defines two points on the graph, which are connected by a straight line. The inverse transform technique applies directly to generating call admission ratio variation, *X*. Recalling the graphical interpretation of the technique, first select adjusted predictive average weighted interference, x. Symbolically this is written as [8]

$$X_1 = F^{-1}(x) \qquad (13)$$

but algebraically, since r is between r_i and r_{i+1}, $X_1 (= \Omega_1)$, is computed by a linear interpolation between x_i and x_{i+1}; that is,

$$\Omega_1 = x_j + \left[\frac{r - r_j}{r_{j+1} - r_j}\right](x_{j+1} - x_j) \qquad (14)$$

when $r = 0.68$, r is between 0.69 and 0.6 ($r_3 = 0.63 < r = 0.68 \le x_4 = 0.69$). Therefore (r -0.63) / (0.69-0.63) is 0.833, so that X_1 will be the distance between 0.36 and 0.48 since r is the way between 0.63 and 0.69; that is,

$$\Omega_1 = 0.32 + \left[\frac{r - 0.63}{0.69 - 0.63}\right](0.48 - 0.32) = 0.453 \qquad (15)$$

Notice that for all r_s between the interval (0.63, 0.69), the value a_4=(0.48-0.32)/(0.69-0.63)=2.667 will be needed to compute X_1 , The value a_4 is the gradient $\Delta r / \Delta x$ of the function $x = F'^{-1}(r)$, which is merely the reflection of the line r=x of the function r=$F'(x)$ of Fig. 2. As a result of it, suppose r=0.68, by the Table II, due to $r_3 = 0.63 < r = 0.68 \le x_4 = 0.69$, r lies in the interval i=4, and X_1 is 0.453. Therefore, we can control the expected number of calls(Ω_{vcar}) being admitted, that is 0.547(1-X_1) [9].

TABLE II. INFORMATION FOR BASIC CALL REJECTION RATIO

Level (i)	Interval for adjusted predictive average weighted interference	Utilization Dist.	Cumulative Utilization (x_i)	Rejecting Dist.	Cumulative Rejecting Dist. (r_i)	Gradient $a_i = \frac{\Delta r_i}{\Delta x_i}$
1	$0 \le x \le 0.50$	0.5	0.5	0	0	-
2	$0.50 < x \le 0.57$	0.07	0.57	0.16	0.16	2.286
3	$0.57 < x \le 0.63$	0.06	0.63	0.16	0.32	2.667
4	$0.63 < x \le 0.69$	0.06	0.69	0.16	0.48	2.667
5	$0.69 < x \le 0.76$	0.07	0.76	0.15	0.63	2.143
6	$0.76 < x \le 0.83$	0.07	0.83	0.16	0.79	2.286
7	$0.83 < x \le 0.89$	0.06	0.89	0.16	0.95	2.667
8	$0.89 < x \le 1.00$	0.11	1.00	0.05	1.00	0.455

C. Concept of Call Admission and Rejection

Assuming that call state keeps in one state in one period, the switching probability between in-adjacent states is very small, so it can be assumed switching just happens between adjacent states. In wireless communication systems, if the call admission scheme of next call is determined according to current call processing condition, and call state is normally in *j*-th state, the following three cases will happen.

- Call admission scheme is adjusted according to (j-1)-th call state in which calls are allowed, in t his case, call efficiency is lost but communicatio n quality can be achieved.
- Call admission scheme is adjusted according to j

 ISBN: 978-1-61208-279-0

-th call state that comes up to maximum boundary to be processed, in this case, no call efficiency is wasted and communication quality can also be achieved.

- Call admission scheme is adjusted according to (j+1)-th call state that exceeds maximum boundary to be processed, in this case, the call cannot satisfy communication quality request, accordingly call rejection scheme adjusted by call type is needed.

D. CAC Scheme

CAC scheme shows the entire controlling process of the call admission situations, which is formed of an interval selection, a call rejecting ratio calculation, gradient calculation through linear interpolation, and call control by the variable call admission ratio (VCAR). The variable call admission ratio can be derived from BCAR by calculating the linear interpolation, which can improve the call admission ratio. According to the variable call admission ratio, the decision of whether or not to accept calls from mobile subscriber is subject. The CAC scheme on base station is as follows:

1). We calculate total interferences of base station by using an matrix of each base station.

2). Adjusted predictive average weighted interference (ρ_{j_apawi}) receives from call admission activation module monitored a weighted interference to activate a heavily loaded situations,

3). The scheme tries to find an interval for adjusted predictive average weighted interference from Table II,

4). If found, $a_j = \Delta r_j / \Delta x_j$ to calculate the BCAR is given by inverse transform,

5). Then, calculate VCAR through linear interpolation based on the calculated BCAR,

6). Finally, call requested from base station is adjusted by variable call admission ratio according to the adjusted predictive average weighted interference value.

If the level of interval for adjusted predictive average weighted interference is 1, state is normal. If it is 2, state is ready. Otherwise each type of call can be accepted or rejected by the level of interval for adjusted predictive average weighted interference.

V. NUMERICAL RESULTS

A. Simulation Assumptions

Here numerical results are given for adaptive CAC. The processing of calls in base stations usually takes place in several consecutive steps, separated in time. When such systems are simulated, the calls are generated pseudo-randomly, for instance with an exponential distribution of inter arrival times. We suppose the number of the maximum subscriber 12,000 if the busy hour traffic per subscriber is 0.06 Erlang, when the total Erlang is 720(12,000 * 0.06). For CAC, the blocking ratio is calculated by using the well-known Erlang B formula. The calls are assumed to have a constant holding time of 100 seconds for busy hour call per subscriber: 2.16 calls (0.06 * 3,600 / 100). Completed calls in busy hour are 25,920 (12,000 * 2.16). For offered calls, it is still possible that maximum base station efficiency (= permitted weighted interference) used for base station is below 95%. From the assumptions mentioned above, we simulated that the traffic is from 12,960 up to 38,880, when the offered load is from 50% (normal call state) to 150% (heavy transient call state) [1]-[2], [10].

B. Performance Analysis

Fig. 3 compares the proposed scheme with the conventional one in terms of weighted interference, where the ordinate represents the weighted interference and the abscissa the time that is determined by call admission activation scheme. Fig. 3 shows the examples of between *without_cacl* and *with_cac* under the call admission state as the weighted interference. The *without_cac* means that the offered load results from the processing of the conventional scheme. The *with_cac* means the results from the processing of scheme. The adjusted predictive average weighted interference increases due to the increase of weighted interference depending on lots of call fail. There is a difference in the figure, which system performance of proposed scheme is higher than conventional one under *with_cac*. It is observed that a hot-spot cell has to reduce the number of calls to maintain the same QoS. The difference between proposed scheme and conventional one can be explained by the reanalysis of weighted interference with calls. The main simulation performance measures are variable call admission ratio for basic call admission ratio.

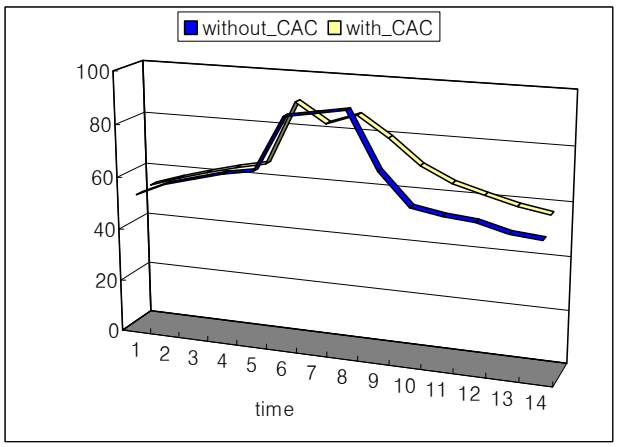


Figure 3. The variance of weighted interference between with and without CAC

Fig. 4 compares the proposed scheme with the conventional one in terms of call admission ratio, where the ordinate represents CAR. Fig. 4 shows examples between for basic call admission ratio (BCAR) and variable call admission ratio (VCAR) under the call

admission state as the weighted interference. VCAR decreases due to the reduction of BCAR depending on the *with_cac*. Especially, in time period-7, note that BCAR decreases to 5%, whereas the proposed scheme is able to keep 20% of VCAR on a level with the previous state(time period-6). The result shows that VCAR does have a higher capacity than that of conventional approach based on BCAR. This phenomenon can be explained by the fixed value of BCAR for the conventional scheme. It is observed that a hot-spot cell has to reduce the number of calls to limit the heavy traffic load as an interference of base stations.

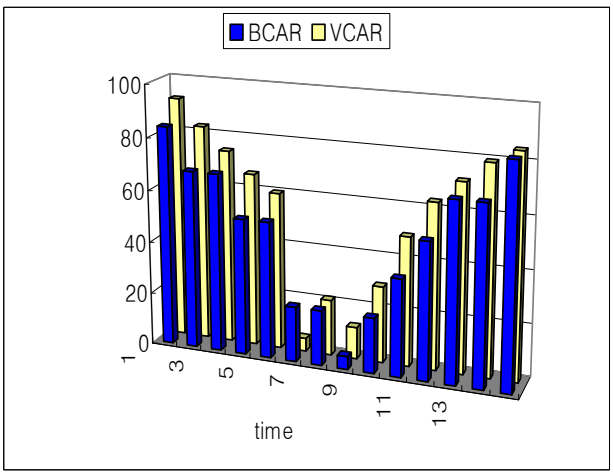


Figure 4. The variance of call admission ratio

VI. CONCLUSION

We proposed new CAC scheme that offers stability of calls in a hot-spot area possibly, and ensures both QoS on call arrival and handover processing. We regard call arrival ratio as a traffic load as interferences, which means that it obtains performance improvement of about 4.3% compared with the conventional approach is not capable of detecting the heavy transient call attempts, whereas our proposed scheme allows system to overcome the problems encountered in the conventional approach regulating calls well. The call arrival, handover, interference, and call success ratio are derived based on the hot-spot environment and traffic model. In conclusion, it is emphasized that in addition to the weighted interference and offered load presented here, congestion control mechanism among base stations must also take into account a variety of technological issues (e.g., a number of admission ratio, an admission/rejection interval, cell radius).

ACKNOWLEDGMENT

This work was supported by the IT R&D program of MKE/KEIT, Republic of Korea (Project No. 10038765, Development of B4G Mobile Communication Technologies for Smart Mobile Services).

REFERENCES

[1] K. S. Lee, "Approximation of the Queue Length Distribution of General Queues," ETRI journal, Vol. 15 No. 3/4, pp. 35-45, 1994.

[2] J. S. Kaufman and A. Kumar, "Traffic CAC in a Fully Distributed Switching Environment," ITC-12, pp. 386-394, 1989.

[3] K. Kim and Y. Han, "A Call Admission Control Scheme for Multi-ratio Traffic Based on Total Received Power," IEICE Transaction on Communications, vol E84-B, pp. 457-463, March 2001.

[4] L. Wang and W. Zhuang, “Call Admission Control for self-similar Data Traffic in Cellular Communications,” GLOBECOM, WC25-2, 2003.

[5] Y. Ishikawa and N. Umeda, “Capacity Design and Performance of CAC in Cellular CDMA Systems,” IEEE Journal on Selected Areas in Communications, Vol. 15, No. 8, pp. 1627-1635, 1997.

[6] Y. N. Han, K. C. Han, and H. G. Bahk, "A CDMA-based Digital Cellular Infrastructure: CDMA Mobile System(CMS)," International Workshop on Multi-dimensional Mobile Communications, 1994.

[7] V. Poxson, and S. Floyd, “Wide area traffic: The failure of Poisson modeling”, IEEE Trans. Networking, Vol. 3, No. 3, pp. 226-244, June 1995.

[8] J. Banks, J. S. Carson II, and B. L. Nelson, Discrete-Event System Simulation, Prentice Hall International Editions, 1996.

[9] S. Catreux, P. F. Driessen, and L. J. Greenstein, “Attainable Throughput of an Interference-Limited Multiple-Input Multiple-Output (MIMO) Cellular System,” IEEE Trans. On Communic., Vol. 49, No. 8, pp. 1307-131, August 2001.

[10] H. Y. Tung, K. F. Tsang, L. T. Lee, K. L. Lam, Y. T. Sun, S. K. R. Kwan, and S. Chan, “On the Handover Performance of a Tri-threshold Bandwidth Reservation CAC Scheme,” ETRI Journal, Volume 29, Number 1, pp. 113-115, February 2007

[11] I. A. Qaimkhani and E. Hossain, “Efficient Silence Suppression and Call Admission Control through Contention-Free Medium Access for VoIP in WiFi Networks,” IEEE Communications Magazine, Vol. 46, Issue 1, pp. 90-99, January 2008.

 ISBN: 978-1-61208-279-0

Design of Optical Wireless IR-UWB Systems for Low Data Rate Applications

Mohammed Al-Olofi, Andreas Waadt, Guido H. Bruck, and Peter Jung
Department of Communication Technologies
University of Duisburg-Essen
Duisburg, Germany
Email: info@kommunikationstechnik.org

***Abstract*—The use of Impulse Radio-Ultra Wide Band (IR-UWB) introduced a low data rate and low power consumption systems, which are suitable for sensor networks. Due to strict regulations of the power transmission and frequency use, the range and usage of IR-UWB remained very limited. To enlarge the range of UWB, this paper presents a system design on how to replace the radio part of the physical layer (PHY) of UWB by free space optics, without changing the major parts of the standard, including frame structure and coding. The performance of such a system on optical links is analyzed in Monte-Carlo simulation. The utilization of optical links will offer new configurations for uplink and downlink. This system will deal with optical noise and optical multipath channel and shows the performances of the new system in presence of Line of Sight (LOS) and diffuse links.**

Keywords- Hybrid Optical/Radio systems; IR-UWB; OWC.

I. INTRODUCTION

The Optical Wireless Communications (OWC) is an alternative technology to radio communications, which suffers from congested frequency bands as the number of mobile users increased significantly. The OWC offer a broad unlicensed free spectrum that enables high data rate, low cost, high speed, and ease of development systems. These advantages make the optical solution attractive for short range communications applications, such as smart homes, smart offices, wireless LANs, and sensor networks.

Currently, the ultra wideband systems are the best technology choice for short range communication, since they offer a large bandwidth (3.1-10.6 GHz), high speed, immune to multipath fading, multi access capabilities, and low cost transceivers. The fractional bandwidth of UWB is defined by FCC as a signal with 20% of its center frequency or 500 MHz bandwidth, when the center frequency is above 6 GHz with a limited power of -41.3 dBm/MHz [1]. The optical devices introduce a low modulation bandwidth, which does not exceed 20 MHz in case of using LED and hence limit the data rate and available broadband spectrum [2]. This limitation is caused by the characteristics of LEDs available in the market, which have a slow response to the feeding current. The 3dB frequency and hence the modulation bandwidth depends on minority carrier lifetime. To solve the limitation of modulation bandwidth problem, we recommend two solutions. We can use either special LEDs with larger modulation bandwidth similar to that used in fiber optical communications or we can use laser diodes, which offer bandwidth in GHz at expense of increasing shadowing effect due to the nature of direct light generated by laser diodes. Nevertheless, a lot of solid state electronics researches focus on optical wireless communication and this gives a sign that we will see high-speed low-cost LEDs in the next decade.

The utilization of optical wireless links in IR-UWB systems will overcome the limitations in emission power and bandwidth and/or introduce a new Radio/Optical IR-UWB system where we can take advantages of both configurations. The optical links do not interfere with the electrical systems and they are not affected by multipath fading. This design will be suitable to operate in sensitive environments like hospitals where the radio communications are prohibited.

In this paper, we have proposed a new optical IR-UWB system design to solve the problem discussed above. The conventional IR-UWB system employing radio antenna at the transmitter and receiver has been modified to be able of transmitting the information data via the optical link. To establish an optical link in our design, both antennae in transmitter and receiver are replaced by LEDs and photodetector respectively. The transmitted pulse is converted into light pulse by regulating the input current of light device to form the same shape of radio pulse. On the receiver side, the power of the received light signal will be detected by photodetector and converted at output to current levels which reconstruct the shape of transmitted pulse. The further steps of detection process are done as in IR-UWB radio systems.

The rest of paper is organized as follows. In Section II, the system design describes the transmitter and receiver is presented. Also in Section II, the optical channel impulse response regarding the environment and SNR are explained. The system simulation is introduced and the results are discussed in Section III. Finally, Section IV concludes the paper.

II. SYSTEM DESIGN

A. Transmitter Design

The transmitter of IR-UWB is modified to be able to transmit in the manner similar to optical pulses as shown in Fig. 1. We employ the Binary Pulse Position Modulation (2-PPM) as presented in [1] combined with intensity modulation.

 ISBN: 978-1-61208-279-0

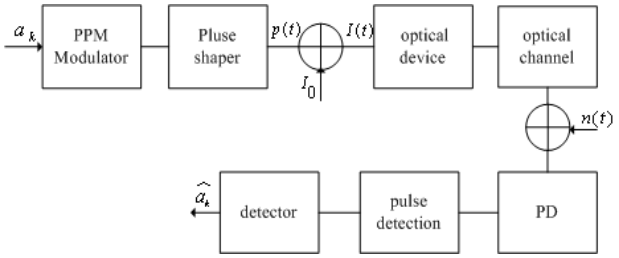


Figure 1. Optical IR-UWB system

.In this modulation scheme, the transmitted pulse $p(t)$ is shifted by ε time shift from the symbol time T_S if the transmitted bit is '1' and introduce no shift when the bit '0'.The modulated signal can be described by

$$x(t) = \sum_{k=-\infty}^{+\infty} p\left(t - kT_s - a_k\varepsilon\right) \tag{1}$$

The Gaussian monocycle pulse $p(t)$ is used to drive the LED by convert it to a set of quantized current levels as shown in Fig. 2. The pulse is quantized to L current levels and every level is mapped to a defined brightness level. The current levels represent the intensity power should be non-negative to ensure that the LED is not reversely biased.

$$I(t) = I_0 + \left|p(t)\right| \tag{2}$$

A dc-bias is chosen to boost the negative part of the radio pulse in order to keep the diode in the 'ON' state and illuminates a 10% of full brightness. The constant forward current (I_0) will keep the diode in 'ON' state even if the pulse time end and hence keep the diode operating in the active region. The modulation bandwidth should be increased since the diode rise and fall time regarding the diode switching operation is minimized.

To create a different emission power for each of current level, we suggest the white-LED lnGaN/GaN to transmit the optical pulses. This white-LED proposed in [5] shows that the emission power in the blue spectrum portion is increased relative to the injected current. Nevertheless, a blue filter at the receiver Front-end should be used to gain the blue spectrum power.

B. Optical Multipath Channel

The optical multipath channel is characterized by an impulse response $h(t)$, which describes the propagation of the optical signal between the transmitter and receiver. The propagation pattern is approximated by lambertian radiation pattern, which state that the light intensity emitted from a source has a cosine dependence on the angle of emission with respect to the surface normal [2,3]. The luminous intensity in angle ϕ is given by

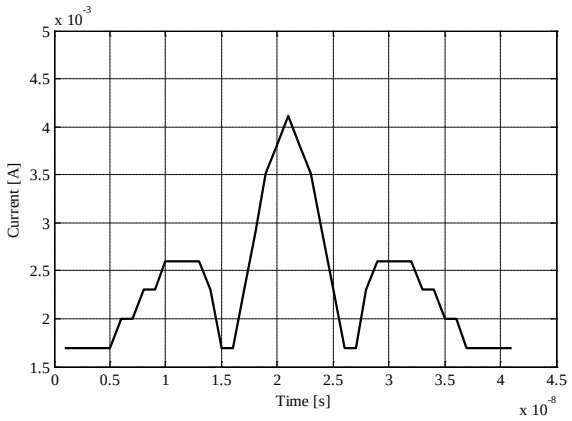


Figure 2. Quantization of the monocycle pulse

$$I(\phi) = I(0)\cos^m(\phi) \tag{3}$$

where $I(0)$ is the center luminous intensity of the LED and ϕ is the angle of irradiance, m is the order of lambertian emission and is given by the semi angle at half illuminance of the LED $\phi_{1/2}$ as

$$m = \frac{-\ln(2)}{\ln(\cos\phi_{1/2})} \tag{4}$$

and the horizontal illuminance I_{hor} at a point (x, y, z) on the working plane is defined as

$$I_{hor}(x, y, z) = \frac{I(0)\cos^m(\phi)}{d^2\cos(\psi)} \tag{5}$$

where d is the distance between the transmitter and receiver and ψ is the angle of incidence.

In an office room environment, the light arrives receiver directly as LOS link or after number of reflections (diffuse link). The impulse response at zero reflection is given as

$$h_{los}(t) = \frac{A_r(m+1)}{2\pi d^2}\cos^m(\phi).T_s(\psi)\cdot g(\psi) \times \cos(\psi)\cdot\delta\left(t - {}^{d}\!/\!_{c}\right), \quad 0 \le \psi \le \psi_{con} \tag{6}$$

where $T_S(\psi)$ is the filter transmission, $g(\psi)$ and ψ_{con} are the concentrator gain and Field Of View (FOV) respectively.

The gain of the optical concentrator at the receiver is defined by

$$g(\psi) = \begin{cases} \dfrac{n^2}{\sin^2\psi_{con}}, & 0 \le \psi \le \psi_{con} \\ 0, & 0 \ge \psi_{con} \end{cases} \tag{7}$$

where *n* is the refractive index.

To model the reflections, every wall is partitioned to a number of small areas, which act as a new lambertian sources when light incident on it. The impulse for the first reflection is given by

$$h_{ref}(t)=\begin{cases}\dfrac{A_r(m+1)}{2(\pi d_1 d_2)^2}\rho A_{wall}\cos^m(\phi_r).\cos(\alpha_{ir}) \\ \times\cos(\beta_{ir})\cdot T_s(\psi)\cdot g(\psi)\cdot\cos(\psi_r)\cdot \\ \times\delta\left(t-\dfrac{d_1+d_2}{c}\right), & 0\le\psi_r\le\psi_{con} \\ 0, & \psi_r\ge\psi_{con}\end{cases} \quad (8)$$

where d_1 and d_2 are the distances between the LED and a reflective point, and between a reflective point and a receiver surface, ρ is the reflectance factor, A_{wall} is a reflective area. The angles α_{ir} and β_{ir} are represent the angle of incidence to a reflective point and the angle of irradiance to a receiver, respectively, ϕ_r and ψ_r are the angle of irradiance from LED to a reflective point and angle of incidence from reflective point to a receiver.

The optical channel is characterized by the room dimensions, reflectance indices of walls and transmitter and receiver orientation. Table I describes the parameters used to simulate the channel impulse response.

The optical impulse response is used to calculate the channel gain, which is important to estimate the influence of channel on the received power. The power contained in a LOS component is larger than the power contained in the first reflection components as shown in Fig. 3 since the long distance and reflection from the surfaces introduce a power loss. The received power can be calculated when the transmitted power was 1 Watt as

$$p_r=\left(p_t\cdot H_{los}(0)+\int_{ref}p_t\cdot H_{ref}(0)\right) \quad (9)$$

Another important feature is the root mean squared (RMS) delay, which describes how much delay added by the channel. A large delay led to intersymbol interference (ISI), which make the detection of transmitted signal complicated. The RMS delay calculated from the channel impulse response as

$$\tau_{RMS}=\sqrt{\frac{\int(t-\tau_0)^2 h^2(t)\,dt}{\int h^2(t)\,dt}} \quad (10)$$

where τ_0 is the delay time and defined as

TABLE I. CHANNEL PARAMETERS

	Parameter	Value
Room	Room size	5×5×3 m^3
	ρ_{wall}	0.8
Transmitter	Location	(2.5, 2.5, 3)
	M	1
	Elevation	-90°
	Azimuth	0°
	Power	1
Receiver	Location	(0.5,1,0)
	A_r	1 cm^2
	FOV	85°
	Elevation	90°
	Azimuth	0°

$$\tau_0=\frac{\int t\cdot h^2(t)\,dt}{\int h^2(t)\,dt} \quad (11)$$

The RMS delay for the simulated channel is 0.49 ns, which define an upper bound for the transmission rate. For the proposed system, the symbol time is less than the RMS delay and hence no equalizer at the receiver side is needed.

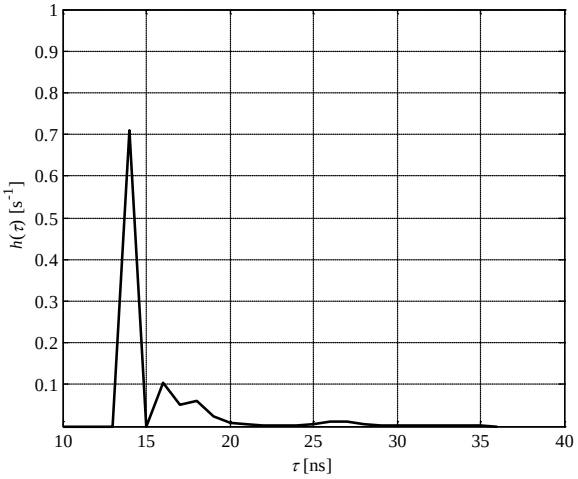


Figure 3. Optical channel impulse response

C. SNR

The SNR in optical system is defined by the received power, photo detector responsivity R [A/W] and noise variance σ^2 as

$$SNR = \frac{(R\, p_r)^2}{\sigma_{sh}^2 + \sigma_{th}^2} \tag{12}$$

In an optical wireless system, the noise is classified into two types namely the shot noise and thermal noise. The shot noise is a time-varying process generated by external light sources like background noise and quantum noise or internal source as intensity radiation, dark noise and excess noise. These sources are independent Poisson random variables and their photoelectron emission follows the distribution of Poisson distribution with mean equal to the sum of the individual processes. The variance of any shot noise process associated with photodetection are represented as

$$\sigma_{sh}^2 = 2\,q\,B\langle i \rangle \tag{13}$$

where q is the electronic charge, B is the equivalent bandwidth, and $\langle i \rangle$ is the mean current generated by $\langle n \rangle$ electron. However, if the photoelectron count is large, the generated signal current probability distribution can be approximated to be Gaussian process [6].

$$p(i) = \frac{1}{\sqrt{2\pi\sigma_{sh}^2}} \exp\left(-\frac{(i - \langle i \rangle)^2}{2\sigma_{sh}^2} \right) \tag{14}$$

Addition to the shot noise, the thermal noise caused by thermal fluctuation of electrons in receiver circuit add a currents, which is Gaussian process has a zero mean and its variance described as

$$\sigma_{th}^2 = \frac{4\kappa T_k B}{R_L} \tag{15}$$

with K is the Boltzmann's constant, T_k is absolute temperature, and R_L is the equivalent resistance. The total generated current probability distribution of thermal noise and shot noise can be represented as

$$p(i) = \frac{1}{\sqrt{2\pi(\sigma_{sh}^2 + \sigma_{th}^2)}} \exp\left(-\frac{(i - \langle i \rangle)^2}{2(\sigma_{sh}^2 + \sigma_{th}^2)} \right) \tag{16}$$

To evaluate the SNR required for a BER, the received power needed to achieve a BER of 10^{-6} is about 18dB.

D. Receiver Design

The receiver of optical system is based mainly on photo-detector employing the direct detection. The area of detector and the orientation play important roll in the receiver design and performance. The photodetector generates an output photocurrent relative to the incident light power pinging on the surface. i.e., the changes produced in intensity modulation at the transmitter are detected by direct detection at the receiver. The photocurrents induced by photodetector form a replica of the transmitted pulse. At the receiver Front-end the received signal is defined as

$$y(t) = R\, I(t) \otimes h(t) + n(t) \tag{17}$$

where $h(t)$ is the optical multipath channel, $I(t)$ is the transmitted signal, and $n(t)$ is the Additive White Gaussian Noise (AWGN). After conversation of optical signal to electrical signal, the correlation between the mask of the transmitted pulse *m(t)* and the received signal *y(t)* is performed as

$$\begin{aligned} m(t) &= I(t - \tau - kT_s) - I(t - \tau - kT_s - \varepsilon) \\ Z &= \int_{\tau}^{T_s + \tau} y(t)\, m(t)\, dt \end{aligned} \tag{18}$$

The detector in (19) compares the power of correlation *Z* to a threshold and decides whether the received bit is '0' or '1' [7].

$$\hat{a} = \begin{cases} Z > 0 & \hat{a} = 0 \\ Z < 0, & \hat{a} = 1 \end{cases} \tag{19}$$

III. SIMULATION RESULTS

The system design represented in Fig. 1 is simulated using MATLAB program. The Monte Carlo simulations were carried out to generate the bit-error-ratio (BER) versus E_b/N_0 figures. The information bits are modulated by 2-PPM modulator with symbol time $T_s = 240\,ns$, sampling frequency $f_c = 1GHz$, and shift time $\varepsilon = 120\,ns$. The mono-cycle pulse width $T_P = 41ns$ quantized to $L = 8$ current levels, which drive the optical transmitter device and each of the current level generates a one of the brightness level.

The optical channel impulse response in Fig. 3 is simulated in a room with dimensions of (5m×5m×3m) and a fixed transmitter and receiver are assumed. The optical pulse is convolved with the channel and added to the noise.

On the receiver side, the photo detector is percepts the incident light and converts it into current. We assume that the detector responsivity R equal to one. The received pulse constructed from current levels is correlated with the mask of transmitted pulse and the peak power is compared

 ISBN: 978-1-61208-279-0

to the threshold in time window. This system is assumed to be synchronized and no equalization stage is performed.

In OWC systems, a unit rectangular pulse is used to transmit the power of modulated binary bits with duration T_s. [8, 9]. In Fig. 4, the same system design was simulated using rectangular pulse and monocycle pulse to compare the BER performances of using the shaped pulse used for wireless system and rectangular pulse used in OWC systems. Fig. 4 shows that the BER for both signals are equal, which is evident that other pulses shape could be used without loss of performances. Although rectangular pulse evaluation is simpler than Gaussian pulse, the later introduces capability for utilizing advantages of a designed UWB radio wireless system communicating on optical link in sensitive environments. Nevertheless, the effects of LEDs nonlinearities and shot noise are expected to disfigurement the transmitted Gaussian pulse at transmitter and receiver front-ends. These effects will be studied experimentally in the future work to find the performance degradation for the proposed design.

Fig. 5 compares the BER performance of the proposed systems operate on LOS optical wireless channel with that on diffuse channel in the absence of LOS link. The direct path between transmitter and receiver delivers higher power than paths reach the photodetector after reflections. This explains why the BER in the presence of LOS channel is lower than that in diffuse channel by ~2dB. Also, the influences of optical wireless channel gain and delay on the proposed system increase the BER by ~4dB compared to the AWGN bound. The use of equalization techniques such as zero forcing or decision feedback equalizer will enhance the system performances because the shape of the transmitted pulse at receiver is better restored.

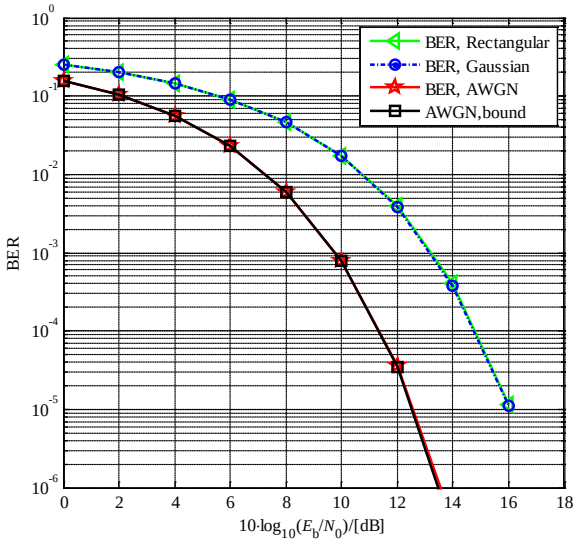


Figure 4. BER of the system with rectangular and Gaussian Pulses.

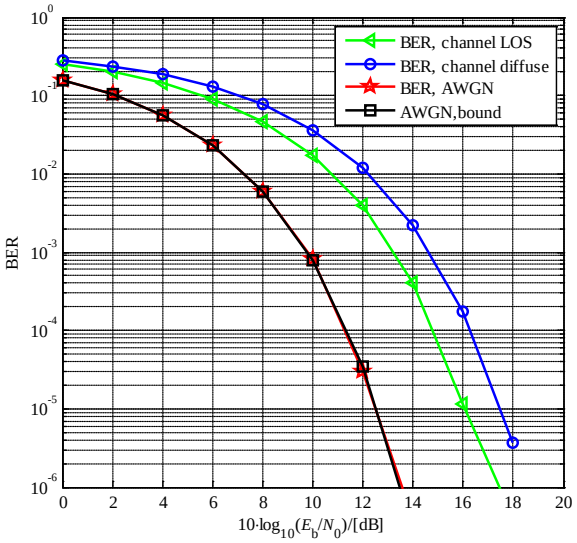


Figure 5. BER of the system with LOS and diffuse link configurations

IV. CONCLUSION AND FUTURE WORKS

In this paper, a new design of optical IR-UWB is introduced to overcome the limitations of power and bandwidth since optical links introduce unlimited bandwidth and no constrains on the emitted power. Our design will be suitable for non radio environments like hospitals or for systems that operate in both optical/radio configurations. The optical wireless link is established between the transmitter and receiver by replacing antennae with optical devices. Also, the radio pulse is quantized and offset is added to define an optical version of the radio Gaussian pulse. The optical pulse is transmitted using an optical device and received by a photodetector. The effects of the optical multipath channel are investigated and the system performance using Monte-Carlo simulations has been obtained and analyzed.

This work will be continued to introduce a better system performance using equalization techniques. Also, the LEDs characteristics effects on pulse shaping process will be investigated in order to achieve experimental results and build a demonstrator for the proposed system.

REFERENCES

[1] IST PULSERS Phase II D3a3.3: LDR-LT Concept Specifications - PHY and MAC Layers, Jul. 2008.

[2] J. R. Barry, J. M. Kahn, W. J. Krause, E. A. Lee, and D. G. Messerschmitt, “Simulation of multipath impulse response for wireless optical channels,” IEEE J. Select. Areas in Commun. vol. 11, no. 3, Apr. 1993, pp. 367–379, doi: 10.1109/49.219552.

 ISBN: 978-1-61208-279-0

[3] H.Q. Nguyen *et al.*, "A MATLAB-based simulation program for indoor visible light communication system," CSNDSP 2010 proceedings, July 2010, pp. 537-541.

[4] T. Komine, M. Nakagawa, "Fundamental analysis for Visible-Light Communication system using LED light," IEEE Transaction on Cunsumer Electronics, Vol. 50, No. 1, February 2004, pp. 100-107, doi: 10.1109/TCE.2004.1277847.

[5] J. Grubor, S. C. J. Lee, K. D. Langer, T. Koonen, and J. W. Walewski, "Wireless high-speed data transmission with phosphorescent white-light LEDs," in Proc. Eur. Conf. Optical Communications (ECOC 2007), Berlin, Germany, Sept. 2007, pp.1-2.

[6] Z. Ghassemlooy, W. Popoola, and S. Rajbhandari, Optical Wireless Communications: System and Channel Modelling, CRC Press, 2012, pp. 66-74.

[7] M. -G. Di Benedetto, Understanding Ultra Wide Band Radio Fundamentals, Pearson Education, 2008, pp.241-252.

[8] M.D Audeh, J.M Kahn, J.R. Barry, "Performance of pulse-position modulation on measured non-directed indoor infrared channels," IEEE Transactions on Communications, vol.44, no.6, Jun 1996, pp.654-659, doi: 10.1109/26.506380.

[9] J. Zhang, "Modulation analysis for outdoors applications of optical wireless communications," International Conference on Communication Technology Proceedings, 2000. WCC - ICCT 2000., vol.2, no, 2000, pp.1483-1487, doi: 10.1109/ICCT.2000.890940.

 ISBN: 978-1-61208-279-0

Using Service Delay for Facilitating Access Point Selection in VANETs

Tin-Yu Wu[1], Wei-Tsong Lee[2], Tsung-Han Lin[2], Wei-Lun Hsu[2],Kai-Lin Cheng[2]

[1]Institute of Computer Science and Information Engineering, National Ilan University, Taiwan, R.O.C.

[2]Department of Electrical Engineering, Tamkang University, Taiwan, R.O.C.

tyw@niu.edu.tw, wtlee@mail.tku.edu.tw, possiblejay@hotmail.com, hsuweiluntw@hotmail.com, mypc_01@yahoo.com.tw

***Abstract*—With the rapid development of wireless mobile networks, VANET (vehicular *ad hoc networks*) that adopt transportation tools as mobile platforms have received great attention. Integrating VANETs with wireless infrastructure to provide high-quality transmission services also has become one of the important research topics. Because terminal devices in vehicular environments are highly mobile, a MN (*mobile node*) will encounter frequent handoffs while accessing wireless network services. However, supposing the chosen handoff AP (*access point*) presents too long service delay, the quality of the Internet, especially real-time services, like VoIP and multimedia streaming, will be greatly influenced. Therefore, by using the packet scheduling architecture for classified service at APs, this paper proposes a handoff scheme based on service delay prediction. According to the scheduling scheme, we can estimate the load and service delay of different access categories of the regional APs. Our proposed scheme allows the MN requesting real-time services to be allocated to the AP with the lower service delay and the chosen AP thus can reduce the service delay for users.**

Keywords- VANET; Wireless network switch; AP selection

I. INTRODUCTION

In recent years, wireless network has been extensively deployed in the environment for users to access network services. Wireless communication, consequently, becomes more and more important in our daily life because users at any locations are able to use wireless network via APs during the moving process. Because of the emergence of abundant real-time network services, the QoS (*quality of service*) of wireless network also becomes significant. By integrating wireless network with telematics, VANETs (*vehicular ad hoc networks*) provide more and more services for vehicles on the roads. Nevertheless, owing to the high mobility of terminal devices in VANETs, handoffs occur frequently while users access wireless network services [1][2]. If the handoff latency is too high, the quality of network services, like VoIP and multimedia streaming, will be greatly influenced [3]. Previous VANET researches mainly focused on the reduction of scan delay during the handoff procedure, but did not consider the delay resulted from the load of the target AP. For example, not only the processing speed of different service types but also the load of different service types in an AP varies. Thus, the existing methods cannot select the most suitable AP to decrease the service delay. For this reason, this paper presents a scheduling scheme to estimate the load and service delay of different service types for the APs in the area and select the optimal AP for each service type for handoff. Our proposed scheme allows the MN requesting real-time services to be allocated to the AP with the lower service delay.

The rest of this paper is structured as follows. Section II introduces the background and related works. Section III explains our proposed scheme, including system architecture, scheduling architecture of APs for different service types, load estimation for APs, service delay prediction for APs, service-oriented plus service delay prediction-based regional load balance, and AP selection flowchart. Section IV uses the simulation to prove that our proposed scheme decreases the service delay for users. Finally, the conclusion and future objective is given in Section V.

II. RELATED WORKS

Every AP on a wireless network is responsible for a specific coverage area. When a MN is leaving the coverage area of the current AP, the MN has to search for the surrounding APs and handoff to the next suitable one. The complete handoff procedure includes scanning, authentication, and re-association. Through the selection mechanism, the MN can determine the most suitable AP, perform the authentication and re-associate with the AP to finish the handoff. All existing handoff mechanisms choose the best AP to handoff based on the RSSI (*received signal strength indicator*) but such mechanisms are not suitable for highly mobile VANETs since the time for a MN to stay with an AP is short and handoffs thus frequently occur. Because MNs choose the target APs by themselves and most of them connect to the network via several specific APs, the service delay increases greatly and consequently affects the quality of network transmissions, especially real-time network services.

Therefore, traditional mobility management mechanisms for Internet and MANET (*mobile ad hoc network*) cannot satisfy the needs of vehicular networks and the performance degrades severely owing to the unique features of vehicular networks [4]. In this paper, we mainly focus on the communication between vehicles and infrastructure of RSU (*roadside unit*) with the attempt to satisfy users' requirements

 ISBN: 978-1-61208-279-0

by improving the handoff according to the characteristics of VANETs.

Up to now, many AP selection schemes have been proposed to determine the best AP to handoff by different load metrics and we will introduce several existing AP selection schemes next. The load metric proposed in [5] is the Maximizing Local Throughput, which is based on the number of MNs connected by each AP and the PER (*packet error rate*). However, the number of MNs only roughly implies the load of each AP because the degree of network utilization of each AP differs and the states of different traffic types alter with the time. In addition, this mechanism depends on the PER very much. When the PER is very low, the load balance cannot be improved efficiently. Another AP selection mechanism presented in [6] chooses the AP in light of the signal strength of APs, the number of MNs and the bit rate. By considering the bit rate between APs and MNs and the number of MNs connected to each AP, [7] estimates the throughput between APs and MNs at the higher rate. In [8], on account of queue congestion that might occur when the network is overloaded, the authors use the frame dropping rates of the current AP as the load metric and regulate the received signal strength from each AP to adjust the coverage area of each AP with the aim of achieving load balancing. Based on the measurements of delay incurred by 802.11 beacon frames, [9] selects the AP with the maximal potential throughput or bandwidth. Due to QoS considerations, [10] proposes the iLB (*integrated Load Balancing*) scheme, which chooses the AP with the lower packet delay and defines a handoff threshold value based on the packet loss rate. The iLB scheme establishes an APC (*AP controller*) to gather related information for the APs periodically and broadcast the information of the adjacent APs to the MNs by beacon frames. For QoS management and congestion control for wireless APS, Tartarelli et al. [11] proposes to establish a Policy Server to differentiate QoS guarantees for different traffic types and users, and to reduce network congestion by traffic distribution. Tüysüz et al. [12] present a novel AP selection algorithm, which uses E-model conversational audio quality factor R to estimate the perceived voice quality as the criterion in selecting the most suitable AP to handoff.

III. PROPOSED LOAD BALANCE SCHEME

In VANETs, MNs connect to APs to access wireless services, but MNs may encounter different levels of service delay because of frequent handoffs. According to the scheduling, we propose to quantize the load of different ACs (*access categories*) in each AP and present the Service Delay Prediction scheme, a service delay-based AP selection mechanism. After the scheme quantizes the service delays of different ACs in each AP, users can select the handoff APs with the lower service delay based on the current services. Such a classified selection scheme distributes the load efficiently in case the throughput is degraded and some APs are overloaded due to load imbalance. Previous AP selection methods chiefly select the best AP to handoff based on various load metrics after the affiliation of MNs. In this paper, we classify the network services, use the service delay of different ACs in each AP as the load metric, and select the most suitable AP that meets the requirements of the MN to decrease the service delay.

A. System Deployment and AP Control Architecture

Based on the AP architecture proposed in [14], this paper divides the vehicular environment into several service areas and APs in each area offer services for local MNs. When several MNs are leaving the current service area, APs in the forward area will provide wireless services to MNs. To guarantee QoS and distribute the load, our scheme gathers users' statuses and APs' load in the forward area to estimate the load of the next handoff AP for MNs to select a suitable AP. As shown in Figure 1, an AP controller is established to periodically gather the area information. Every AP in the area regularly uploads its own information and the conditions of the current MNs in the area to the APC. AP information includes the current channel, location and data amount of each AC that is waiting to be processed. MN information records the traffic flow of all ACs that users are accessing in the area. Supposing a MN is accessing several kinds of services simultaneously, the AC with the highest use frequency is regarded as the representative category. According to the information, the APC estimates the loading states and waiting delay of each AC in APs in the forward area and sends the computed result to the MNs in the current area through the current APs.

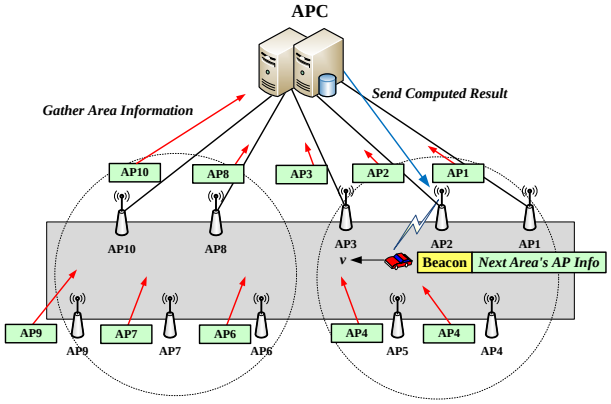


Figure 1. System architecture.

B. QoS for Service Classification

For APs to support QoS-guaranteed service classification, when packets of various traffic types arrive at APs in the vehicular network, the classifier of each AP maps the packets to the ends of the corresponding FIFO (*First In First Out*) queues and the packets in the queues wait to be delivered. Packets in the output queue of each category will be scheduled by the packet scheduler. By the weighted round robin module, the scheduler transmits the packets from the high-priority queues first. Queues of different ACs have

 ISBN: 978-1-61208-279-0

different priority values and thus the number of packets will be different in each round- robin cycle. In every round-robin cycle, the packet number of every queue is fixed and proportional to the weight. As defined in IEEE 802.11e, packets in vehicular environment include four access categories: AC_0=AC_VO, AC_1=AC_VI, AC_2= AC_BE, AC_3= AC_BK and each AP has four output queues to store the packets of VO (*Voice*), VI (*Video*), BE (*Best Effort*) and BK (*Background*). In this paper, the packet size of the four categories is the same and the time for an AP to deliver packets is set to T (Delivery Time Unit). Every AC has its own weight: W_{VO},W_{VI},W_{BE} and W_{BK}, which means the packet number that can be delivered in each round-robin cycle. Packets are sent in order from the highest-priority VO, VI, BE to BK. The total deliver time of packets from four queues in every complete round-robin cycle can be denoted by Equation 1:

$$\sum_{i=0}^{3} W_{AC_i} \times T \qquad (1)$$

Queues with the higher weights can deliver more packets in each round-robin cycle and the processing speed for different queues differs. The scheduler transmits the data of four ACs in order in each round-robin cycle until the packets of each AC are fully delivered.

C. Load Estimation for APs

In the previous section, we have introduced the service-classified packet scheduling architecture for APs. The packet number of each AC that is waiting to be delivered in the AP is set to NNN. Based on the parameters of priority scheduling for the router in [15], we set the weight values of four ACs (W_{VO},W_{VI},W_{BE},W_{BK}) to (8,4,2,1). MU_{AC_i} means the number of round-robin cycles for all the packets in the queues waiting to be transmitted completely and to serve the MNs. Supposing the number of round-robin cycles of two ACs at an AP is the same, the waiting delay for users who access the two ACs will be similar, which implies the loading states of the two ACs are the same. On the contrary, the bigger difference between the numbers of round-robin cycles refers to not only different loading states of the two ACs but also their waiting delay. The number of round-robin cycles displays the comparative busyness of the ACs and the number of rounds that users have to wait.

$$\mu = \frac{\sum_{n=1}^{n} MU_n}{n} = \frac{\sum_{I=0}^{3} MU_{AC_i}}{4} \qquad (2)$$

Equation 2 uses μ to represent MU_{AC_i}, the average number of round-robin cycles of each AC at an AP. In this paper, the value of μ is regarded as the load of an AP. The higher value of μ means the larger average number of round-robin cycles of the ACs at an AP. Therefore, the data that waits to be delivered must be processed by more rounds and the load of the AP is comparatively heavy.

D. Service Delay Prediction for Different Access Categories

Because we use a round-robin algorithm to schedule packets from several output queues, the amount of time needed to process one queue is directly influenced by the amount of data in the other queues. The packet scheduler transmits the packets in the queues in order. When a MN chooses the ACs on an AP, different amount of data in the queues results in different levels of service delay. Figure 2 displays the flowchart for computing service delay of different ACs.

Several basic parameters are defined as follows:

- N_{AC_i} : number of the AC packets waiting for transmission at an AP
- $Total$: AP's service delay caused by the cumulative number of packets sent before serving users
- W_{AC_i} : number of the AC packets that can be transmitted during a round-robin cycle
- MU_{AC_i} : number of round-robin cycles for the AC queue to wait for processing the MN's data
- T : delivery time unit for the AP to transmit packets
- I : serial number of the target AC { 0=VO , 1=VI , 2=BE , 3=BK }

Through the APC, we know the states of the four AC queues in the APs in the forward area, estimate the service delay of the AC queues, and obtains the service delay of each AC while the MN determines the best AP to handoff. Different levels of service delay is required for different service types. Compared with previous load estimation methods, our proposed scheme measures the service delay that users may encounter after choosing the target AP and thus reveals the loading states that users actually experience more. The APC gathers the information of the forward area and obtains μ, the current loading states of the APs, and service delay of each AC according to N_{AC_i}, the data amount of the ACs.

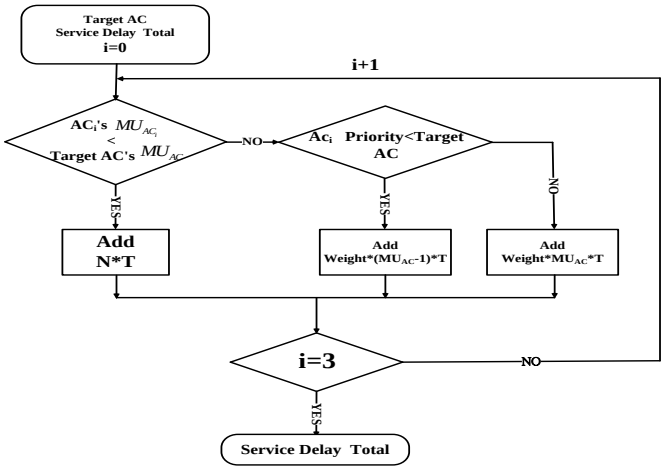


Figure 2. Flowchart for the computation of service delay.

E. *Service-oriented plus Service Delay Prediction-based Regional Load Balance*

This paper presents a service-oriented scheme, which selects the best AP to handoff according to the service delay of all APs. As shown in Figure 3, the vehicle MN1 can choose AP1 or AP2 for the handoff. Because different ACs have different advantages, the processing speed of the four AC queues in each AP is also different and so are the queue lengths. Thus, before MN1 makes the decision, several considerations must be made. As for the simple circumstance, MN1 demands VI service currently, keeps uploading a fixed number of packets to its designated AP, and will choose the best candidate AP to handoff. Figure 3 shows that AP2 is lighter loaded than AP1 but the number of VI packets to be delivered in AP2 is more than AP1. Different queue lengths of the four ACs influence the service delay. For this reason, our proposed scheme considers both the service delay of each AC and the loading states of the APs to guarantee the QoS of each AC and to achieve the regional load balancing. We aim to select an AP with the lower service delay of a specific AC and the less average round level for the handoff.

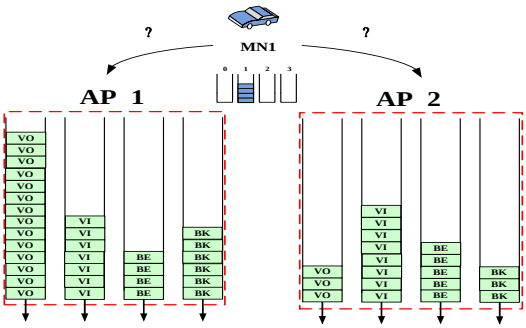


Figure 3. Service-oriented AP selection.

By our scheme, we can get the service delay of each AC on each AP in the forward area. Since AC VO clients cannot endure too long delay, service delay is significant to AC VO and an AP with the lower service delay will be necessary for the handoff procedure. Also, AC VO traffic has low bandwidth requirement. On the other hand, with the design of the buffer, AC VI often can endure longer delay than AC VO, but the data generation rate of AC VI is higher and thus demands higher bandwidth requirement than AC VO. According to ITU-T, G.1010 [13], based on users' tolerance of real-time services, we set the threshold values for service delay, Th_1 and Th_2, and classify the service delay of the ACs on each AP into three levels, as displayed in Table 1.

TABLE I. SERVICE DELAY LEVELS

Service Delay Level	
Level 1	Service Delay<Th_1
Level 2	Th_1<Service Delay<Th_2
Level 3	Th_2<Service Delay

In the APC, each AC has a table of service delay levels to level the APs and APs in the forward area are divided into three levels according to the table. Level 1 means the service delay that is lower than Th_1 and is imperceptible to users. Level 2 means the service delay ranging between Th_1~Th_2 that is perceptible but tolerable to users. The service delay that is higher than Th_2 and intolerable to users is leveled as Level 3. Therefore, four ACs are leveled into different service delay levels based on the same thresholds.

F. *Service Delay Prediction AP Selection Scheme*

The APC gathers the states of each AP in each area, calculates the service delay of each AP, and lists each AP in the target AC's service delay level table. After the APs broadcast the service delay level table of the APs in the forward area, users can examine the table of the target AC and select the AP with the minimum service delay level and the least average round level for the handoff. By the load estimation, our service delay prediction scheme allows the MNs to select the APs that satisfy the service delay of the target AC and choose the handoff AP with the minimum service delay level and the minimum average round level. Instead of choosing the AP with the minimum waiting time, we choose the AP with the lightest load so that the MN can distribute the load efficiently while selecting the AP to handoff. To satisfy QoS guarantee and achieve load balancing, our scheme maps users to different ACs, considers the APs with the minimum service delay level first, and selects the AP with the minimum average round level.

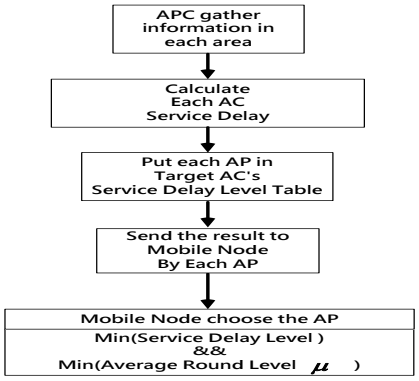


Figure 4. Mobile Node Handoff Architecture.

IV. PERFORMANCE ANALYSIS

This section will simulate and analyze our proposed scheme. Section 4.1 investigates the average service delay variation of four ACs under different traffic flow. Matlab is used as the simulation tool and related steps and parameters are described in the following.

By referring to [14], we group the APs in the areas and choose the best APs to handoff when MNs move from one area to the other. Figure 5 presents the simulation scenario, in which each area is constructed by a fixed number of APs. The overlapping range of the two areas is assumed to be

 ISBN: 978-1-61208-279-0

50M. The average data generation rate of each AC is set based on ITU-T, G.1010[13] and the weight values of the four ACs are set to (8,4,2,1). Compared with (4,3,2,1), our assumed setting conforms to the service delay of the four ACs more under the same load. To choose the most suitable APs to handoff, we predict the service delay that MNs might encounter while moving into the next area. Moreover, to avoid too long service delay affecting the QoS of real-time services and to maintain the load balancing of the next area, we classify the APs and MNs according to the service types and use the average round level and service delay as the load metrics. In the following simulation, with 5 APs in the forward area, we examine the service delay variation of the four ACs when the amount of data increases. The rest parameters are listed in Table 2.

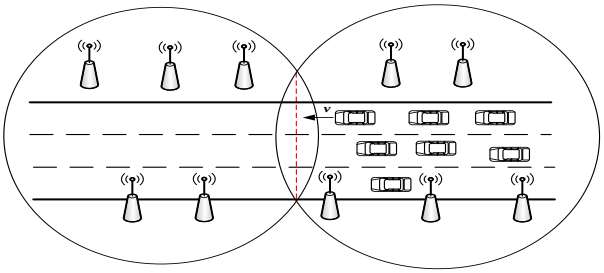


Figure 5. Simulation Scenario.

TABLE II. SIMULATION PARAMETERS

Parameters	Value
Packet Size	1024 bytes
AP Service Rate	100MBps
AP transmission Radius	500m
Overlap range	50m
Mobile Node Speed	100 Km/h
Access Category	VO,VI,BE,BK
(W_{VO},W_{VI},W_{BE},W_{BK})	(8,4,2,1)
Service Upload Rate	8~50 packets/sec
Offered Load(MB)	5MB~30MB
Th_1	150ms
Th_2	400ms

4.1 Analysis of Service Delay

In the scenario, there are 5 APs in the forward area. When the MNs' load during the handoff procedure increases, we use different load metrics and compare the average delay variation of the four ACs under different traffic flow conditions. The results are shown in Figure 6~8. The load metrics include: (1) service delay, proposed in this paper that calculates the service delay of four ACs in each AP and classifies the APs, (2) load balance, proposed in [10] that selects the best AP to handoff according to the packet delay metric, and (3) RSSI. Because Service Delay Prediction scheme levels the APs in the forward area based on service delay, Figure 6 reveals that our method allows the MNs requesting real-time services to have lower service delay. Compared with the packet delay-based scheme presented in [10] and the RSSI-based scheme, considering that different ACs require different service delay, our Service Delay Prediction scheme classifies the APs according to the services to the MNs, classifies the MNs, and chooses the AP with the lowest service delay levels for the handoff.

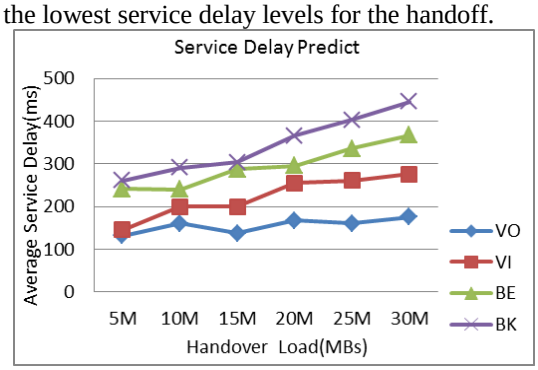


Figure 6. Simulation Scenario. Average delay variation of four access categories (Service Delay Prediction).

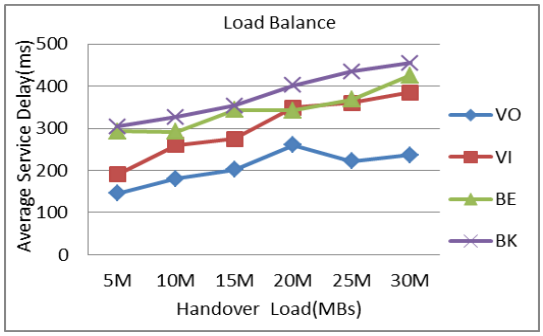


Figure 7. Average delay variation of four access categories (Load Balance).

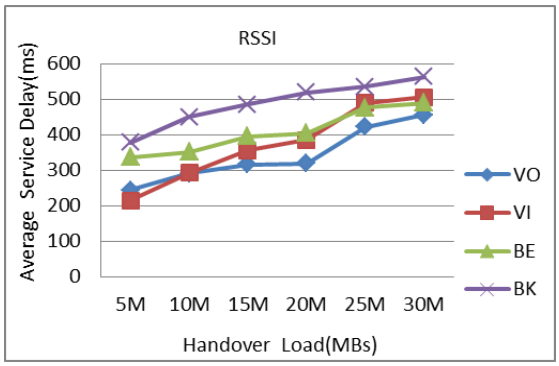


Figure 8. Average delay variation of four access categories (RSSI).

To satisfy QoS guarantee and achieve load balancing, the AP with the minimum average round level is chosen for the handoff. Figure 7 shows that the packet delay-based AP selection scheme proposed in [10] does not consider service

 ISBN: 978-1-61208-279-0

delay caused by different busyness of different ACs in each AP and thus the MNs accessing real-time services cannot have the lower service delay. To select the handoff AP based on the RSSI, Figure 8 reveals that the handoff APs chosen by the MNs will be closer to the previous area and cannot satisfy the service delay for the ACs. Therefore, the service delay of the four ACs is obviously longer than the other two schemes.

4.2 *Analysis of AP Load Balance in the Area*

μ , the average number of round-robin cycles of APs, is taken as the balance index in our proposed method, as shown in Equation (3):

$$\beta(x) = \frac{\left(\sum \mu\right)^2}{N \times \sum \mu^2} \quad (3)$$

To compare packet delay-based load balancing scheme [10], RSSI-based scheme and our Service Delay Prediction scheme, we simulate two situations, in which there are 4 and 5 APs in the area, respectively, to examine the load distribution of different ACs by each scheme in different handover load when the MNs are in the overlapping area. The simulation results displayed in Figure 9 and 10 reveal that by using our Service Delay Prediction scheme, the load balancing method in [14] enables the system to reach load balance, which is very close to 1. As for the RSSI-based scheme, the balance index is low and becomes unstable with the increase of handover load. Because of the RSSI, some APs nearer the previous area are easily selected by the MNs without considering the current load of the APs and the load thus cannot be distributed to the regional APs. The load balancing method presented in [14] chooses the handover AP according to the busyness of AP and thus performs better than the RSSI-based scheme in load distribution. Nevertheless, the load of different ACs in each AP is not considered and therefore the general load dispersion is below the Service Delay Prediction scheme. Our Service Delay Prediction scheme chooses the suitable AP for the MNs moving to the forward area according to the service delay of different ACs in the APs, thus making good use of resources of each APs in the forward area.

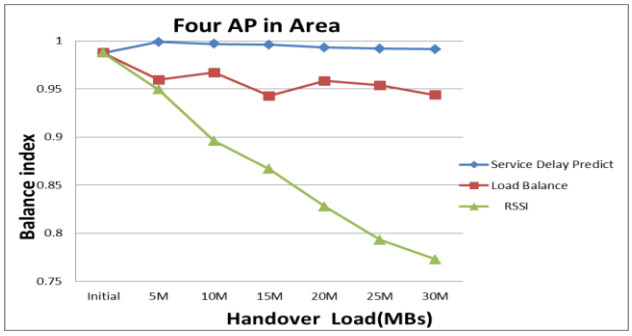


Figure 9. Load Balance Analysis (Four APs in the area)

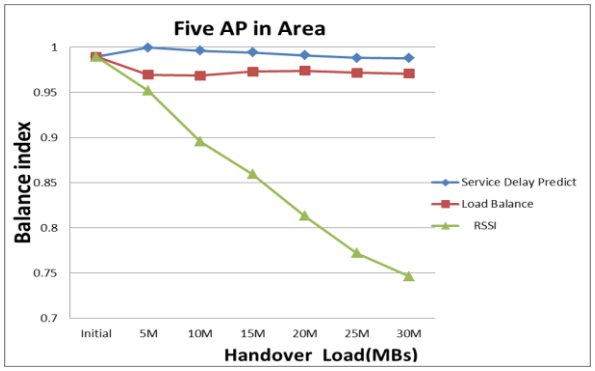


Figure 10. Load Balance Analysis (Five APs in the area)

V. CONCLUSION

Based on the service delay of the APs in the forward area, this paper proposes an access point selection scheme for VANETs, in which an APC (Access Point controller) is established to gather the information of APs and users in each area and compute the service delay of the APs in the forward area. The computed result is sent to MNs by the regional APs. A QoS-guaranteed handoff AP not only reduces the service delay of real-time services in each AP but also maintains the load balancing among the regional APs and the simulation result proves that our proposed method achieves the above-mentioned goals. In the future, we aim to include users statuses under frequent handoffs during long-term movement to enhance the regional load-balancing capacity in the area. Also, to classify the services for further discussions will be more pertinent to actual utilization.

ACKNOWLEDGMENT

This study was supported by the National Science Council, Taiwan, R.O.C., under grant no. NSC 101-2219-E-197-003.

REFERENCES

[1] D. Kwak, J. Mo, and M. Kang, "Investigation of Handoffs for IEEE 802.11 Networks in Vehicular Environments," in Proceedings of the First International Conference on Ubiquitous and Future Networks, Hong Kong, China, June 2009, pp. 89-94.

[2] N. Choi, S. Choi, and Y. Seokt, et al., "A Solicitation-based IEEE 802.11p MAC Protocol for Roadside to Vehicular Networks," Mobile Networking for Vehicular Environments, 2007, pp. 91-96.

[3] Y. A. Powar and V. Apte, "Improving the IEEE 802.11 MAC Layer Handoff Latency to Support Multimedia Traffic," Wireless Communications and Networking Conference (WCNC 2009), 2009, pp. 1-6.

[4] K. Zhu, D. Niyato, P. Wang, E. Hossain, and D. I. Kim, "Mobility and handoff management in vehicular networks: a survey," Wireless Communications and Mobile Computing, Oct. 2009.

[5] Y. FUKUDA, and Y. OIE, "Decentralized Access Point Selection Architecture for Wireless LANs Deployability and Robustness," Vehicular Technology Conference, vol. 2, 2004, pp. 1103-1107.

[6] Kuang-Hui Chi, and Li-Hsing Yen, "Load balancing for Non-homogeneous IEEE 802.11 networks using association control," work in progress.

 ISBN: 978-1-61208-279-0

[7] M. Heusse, F. Rousseau, G. berger-Sabbatel, and A. Duda, "Performance anomaly of 802.11b," in Proc. Infocom 2003, San Francisco, CA, USA, 2003, pp. 836-843.

[8] O. Brickley, S. Rea, and D. Pesch, "Load Balancing for QoS Enhancement in IEEE 802.11e WLANs Using Cell Breathing Techniques," Proc. IFIP Mobile and Wireless Communication Networks Conf., Int'l Federation for Information Processing, 2005, www.aws.cit.ie/personnel/ Papers/Paper268.pdf.

[9] S. Vasudevan et al., "Facilitating Access Point Selection in IEEE 802.11 Wireless Networks," Proc. Internet Measurement Conf., Usenix Assoc., 2005, pp. 293–298.

[10] E. H. Ong and J.Y. Khan, "An Integrated Load Balancing Scheme for Future Wireless Networks," International Symposium on Wireless Pervasive Computing (ISWPC 2009), 2009, pp. 1-6.

[11] S. Tartarelli and G. Nunzi, "QoS Management and Congestion Control in Wireless Hotspots," Network Operations and Management Symposium (NOMS 2006), 2006, pp. 95-105.

[12] Tüysüz M.F., and Mantar, H.A., "Access point selection for improving the voice quality and overall throughput in wireless LANs," 2010 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), pp. 165-169, 23-25 Sept. 2010 .

[13] ITU-T, "G.1010: End-user multimedia QoS categories," 2001.

[14] Tin-Yu Wu, Wei-Tsong Lee, Fong-Hao Liu, Hung-Lin Chan, and Tsung-Han Lin, "An Efficient Pre-scanning Scheme for Handoff in Cooperative Vehicular Networks", 22nd Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (IEEE PIMRC 2011) , Toronto, Canada, September 11-14, 2011.

[15] Cisco,Catalyst 2948G-L3 and Catalyst 4908G-L3 Software Feature and Configuration Guide

[16] Tin-Yu Wu and Wei-Fang Weng, "Reducing handoff delay of wireless access in vehicular environments by artificial neural network-based geographical fingerprint", IET Communications, Vol. 5, Issue 4, pp. 542-553, March 2011.

[17] Tin-Yu Wu, Yan-Bo Wang, and Wei-Tsong Lee, "Mixing Greedy and Predictive approaches to Improve Geographic Routing for VANET", Wireless Communications and Mobile Computing (WCMC), Vol. 12, Issue 4, pp. 367-378, March 2012.

[18] Tin-Yu Wu, S. Guizani, and Wei-Tsong Lee, and Kuo-Hung Liao, "Improving RSU service time by Distributed Sorting Mechanism", Ad Hoc Networks, Vol. 10, Issue 2, pp. 212-221, March 2012.

 ISBN: 978-1-61208-279-0

Application of DFT Spreading to OFDM Based WLAN for Energy Efficiency Improvement

Masahiro UMEHIRA, Takuya NISHIZAWA and Shigeki TAKEDA
Graduate School of Science and Engineering
Ibaraki University
Hitachi, Japan
umehira@mx.ibaraki.ac.jp

Abstract— **This paper describes an application of Discrete Fourier Transform (DFT) spreading to Orthogonal Frequency Division Multiplexing (OFDM) based wireless LAN (WLAN) to reduce Peak to Average Power Ratio (PAPR) for energy efficiency improvement and to maintain robustness to DC offset error for cost-effective hardware implementation. We call our proposed scheme null DC sub-carrier DFT spreading OFDM, where the DC sub-carrier is made null by splitting the spectrum in the frequency domain after DFT spreading of the modulated signals. The computer simulation results confirm that BER performance is not degraded due to DC offset error at the transmitter and/or the receiver like OFDM with null DC sub-carrier and its PAPR is lower than OFDM, and is almost the same as the conventional DFT spreading OFDM.**

Keywords-DFT spreading; OFDM; Null; DC sub-carrier; PAPR

I. INTRODUCTION

Wireless local area networks (WLANs) have been widely accepted as a means of broadband wireless access to Internet and have been deployed in various environments such as home, campus and office. According to the increasing demand for mobile/nomadic access to the Internet services, hot spot services based on IEEE802.11a/g/n WLAN standards using 2.4GHz and 5GHz bands are becoming more and more popular in these days to off-load the cellular traffic to WLANs, since most of all mobile personal computers (PCs) and smart phone devices have WLAN access capability. In order to increase the bit rate and throughput of WLAN, IEEE802.11ac/ad are being standardized aiming at multi Gbps broadband access and IEEE802.11ah using UHF band such as 700MHz and 900MHz bands is also being standardized for cellular traffic offload applications and wireless sensor network applications [1]-[3].

As well known, orthogonal frequency division multiplexing (OFDM) is adopted in IEEE802.11a/g/n and will be used in IEEE802.11ac/ad as well due to its robustness to the severe frequency selective fading in mobile environments. OFDM shows excellent transmission performance under severe multi-path fading in mobile environments however the weak point of OFDM is its higher peak to average power ration (PAPR) than that of the conventional single carrier modulation scheme. This high PAPR can be a burden for battery-operated mobile terminal implementation because large output back-off (OBO) required in high power amplifier (HPA) could result in large power consumption. Therefore, IEEE802.11ad specification has an alternative PHY of single carrier modulation in addition to OFDM [1]. For the same reasons, discrete Fourier transform (DFT) spreading OFDM based single carrier-frequency division multiple access (SC-FDMA) has been adopted for the uplink transmission, i.e., from the mobile terminal to the base station, in 3GPP-LTE [4]. As DFT spreading OFDM (DFTs-OFDM herein after) is essentially a single carrier modulation based block transmission scheme, lower PAPR than OFDM can be achieved [5]. In addition, its power spectrum is as compact as that of OFDM. Furthermore, its robustness to the frequency selective fading is also equivalent to OFDM since cyclic prefix (CP) is introduced to avoid the inter-symbol interference due to frequency selective fading [6].

On the other hand, OFDM has another advantage of its robustness to DC offset error for cost-effective hardware implementation, i.e., precise DC offset adjustment is not required at modulator and demodulator since DC sub-carrier is made null in the OFDM based WLAN standards. For example, IEEE802.11a standard uses 52 sub-carriers based on 64 point FFT/IFFT, where DC sub-carrier is not used because BER performance of the DC sub-carrier can be significantly degraded due to DC offset error between D/A converter and I/Q modulator and/or that between I/Q demodulator and A/D converter. Therefore, null DC sub-carrier in OFDM is an important feature of OFDM for cost-effective hardware implementation in WLAN. As mentioned above, DFTs-OFDM is a promising solution to reduce PAPR, however we need to maintain the robustness to DC offset error when we apply DFT spreading to OFDM based WLAN.

In order to solve the above-mentioned problem for the application of DFT spreading to OFDM based WLAN, this paper proposes spectrum splitting after DFT spreading to make DC subcarrier null. In this paper, this new type of DFTs-OFDM is called a null DC sub-carrier DFTs-OFDM, which can achieve both advantages of low PAPR and robustness to DC offset error.

This paper is organized as follows: Section II describes the principle of the proposed null DC sub-carrier DFTs-OFDM (NDCS-DFTs-OFDM, hereinafter), Section III shows the performance evaluation results of NDCS-DFTs-

 ISBN: 978-1-61208-279-0

OFDM by computer simulation, for example PAPR, BER performance and Error Vector Magnitude (EVM) degradation due to DC offset error. They are compared with those of conventional DFTs-OFDM and OFDM. Finally, Section IV concludes this paper.

II. PRINCIPLE OF NDCS-DFTS-OFDM

A. Configulation of NDCS-DFTs-OFDM

The block diagram of the proposed NDCS-DFTs-OFDM is shown in Fig. 1, where Fig.1 (a) shows the transmitter side and Fig.1 (b) shows the receiver side.

As shown in Fig.1 (a), the input data is converted from serial data to M symbols of parallel data, by which M sub-carriers are modulated. M points DFT spreading (pre-coding) is performed for M modulated sub-carriers. Then, M points of DFT spread modulated signals are fed to N points IFFT processor to generate OFDM signals, where N is set at power of two to employ FFT algorithm for reducing signal processing complexity. We assume M is an even number and $M<N$. In the input of IFFT, M point data are divided into two groups and $M/2$ point data are fed to the upper frequency part of IFFT processor input and the other $M/2$ point data are fed to the lower frequency part of IFFT processor input. DC component of the modulated signals is not transmitted by DC sub-carrier, but another sub-carrier. DC sub-carrier with frequency=0 is not used and null data is set at DC sub-carrier in the OFDM signals to avoid the BER performance degradation due to DC offset error. As the DFT spread modulated signals are divided to two parts, PAPR is expected to be slightly larger than that of the conventional DFTs-OFDM due to the insertion of null at DC sub-carrier, but it will be much smaller than that of OFDM. As DC sub-carrier is made null, the modulated signals are not affected by DC offset error between D/A converters and I/Q modulator.

Fig.2 (b) shows the block diagram of the receiver side of NDCS-DFTs-OFDM, where the output of the I/Q demodulator output is fed to A/D converters and CP is removed. As the received signal has no DC component, DC offset error between I/Q demodulator and A/D converters does not affect BER performance. This feature is essentially the same as the OFDM, which does not use DC sub-carrier. After A/D conversion, the received signals are fed to N point FFT and converted to frequency domain signals, where $M/2$ point data at the upper frequency part and the other $M/2$ point data at the lower frequency part are combined to the original modulated signals after DFT spreading at the transmitter, where DC sub-carrier is discarded. M point DFT spreading OFDM signals are fed to M point IDFT processor after frequency domain equalization (FDE). The output of the IDFT processor is demodulated on a sub-carrier by sub-carrier basis.

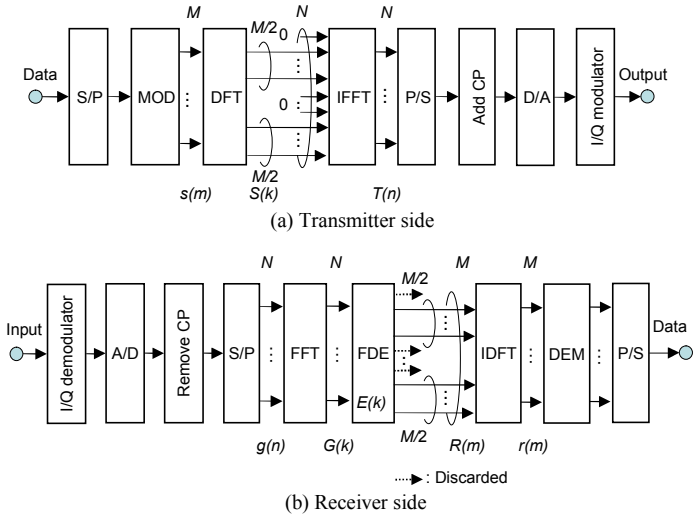


Figure 1. Block diagram of Null DC sub-carrier DFT spreading OFDM scheme (NDCS-DFTs-OFDM).

B. Mathematical expressions of NDCS-DFTs-OFDM

Mathematical expressions of the proposed NDCS-DFTs-OFDM based on the block diagram shown in Fig. 1 are given here.

Let us consider the transmitter side first. We assume that the size of DFT and IDFT is M where M is an even number. Supposing that the complex envelop of the modulated signal is $s(n)$ (n=0~M-1), discrete Fourier transform of $s(n)$, $S(k)$ (k=0~M-1) is given by

$$S(k)=\sum_{n=0}^{M-1}s(n)e^{-j\frac{2\pi}{M}nk}. \quad (1)$$

The matrix expression of equation (1) is given as follows:

$$\begin{bmatrix} S(0) \\ \vdots \\ S(M-1) \end{bmatrix} = \boldsymbol{D}_M \begin{bmatrix} s(0) \\ \vdots \\ s(M-1) \end{bmatrix}, \quad (2)$$

where $\boldsymbol{D}_M$ is M x M square matrix and is given by

$$\boldsymbol{D}_M = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{-j\frac{2\pi\times1\times1}{M}} & \cdots & e^{-j\frac{2\pi\times1\times(M-1)}{M}} \\ \vdots & \vdots & & \vdots \\ 1 & e^{-j\frac{2\pi(M-1)\times1}{M}} & \cdots & e^{-j\frac{2\pi(M-1)\times(M-1)}{M}} \end{bmatrix}, \quad (3)$$

Note that D_M performs M point DFT spreading to convert the time domain signal, $s(n)$ (n=0~M-1) to the frequency domain signal, $S(k)$ (k=0~M-1). $S(0)$~$S(M$-1$)$ are fed to N point IFFT processor, where N is power of two. IFFT converts the frequency domain signals, $S(0)$~$S(M$-1$)$ to the time domain signals, $T(n)$ (n=0~N-1), as shown below:

$$\begin{bmatrix} T(0) \\ \vdots \\ T(N-1) \end{bmatrix} = \frac{1}{N} \boldsymbol{D}_N{}^H \begin{bmatrix} 0 \\ S(0) \\ \vdots \\ S(M/2-1) \\ 0 \\ \vdots \\ 0 \\ S(M/2) \\ \vdots \\ S(M-1) \end{bmatrix}, \quad (4)$$

where $(\ )^H$ is the complex conjugate of the transpose matrix and $\boldsymbol{D}_N$ is N x N square matrix, which is given by:

$$\boldsymbol{D}_N = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{-j\frac{2\pi\times1\times1}{N}} & \cdots & e^{-j\frac{2\pi\times1\times(N-1)}{N}} \\ \vdots & \vdots & & \vdots \\ 1 & e^{-j\frac{2\pi(N-1)\times1}{N}} & \cdots & e^{-j\frac{2\pi(N-1)\times(N-1)}{N}} \end{bmatrix}. \quad (5)$$

The multiplication of $\boldsymbol{D}_N{}^H$ means IFFT processing. After IFFT, CP is added and transmitted like the conventional DFTs-OFDM.

Note that zero is inserted at the first row of the input of FFT processor as shown in equation (4). This means the DC sub-carrier is made null. The frequency domain signals are split into two parts, i.e., $S(0)$~$S(M/2-1)$ and $S(M/2)$~$S(M-1)$. Therefore, the output of IFFT processor is not a pure single carrier modulation signal and DC component, $S(0)$ is transmitted via sub-carrier with frequency=1/N instead of frequency=0. Thus, PAPR could be slightly larger than that of the conventional DFTs-OFDM.

Let us consider the receiver side next. Supposing that the complex envelop of the received signal at the FFT processor input is $g(n)$ (n=0~N-1), the output of FFT processor, $G(k)$ (k=0~N-1) is given by:

$$\begin{bmatrix} G(0) \\ \vdots \\ G(N-1) \end{bmatrix} = \boldsymbol{D}_N \begin{bmatrix} g(0) \\ \vdots \\ g(N-1) \end{bmatrix}, \quad (6)$$

where $\boldsymbol{D}_N$ is given by equation (5). As $g(n)$ suffers frequency selective fading in mobile environments, frequency domain equalization (FDE) is performed to the frequency domain signals, $G(k)$. Supposing that $E(k)$ (k=0~N-1) is the equalization coefficient for linear FDE, the output of FDE, $R(k)$ (k=0~N-1) is given by:

$$\begin{bmatrix} R(0) \\ \vdots \\ R(N-1) \end{bmatrix} = \begin{bmatrix} E(0)\cdot G(0) \\ \vdots \\ E(N-1)\cdot G(N-1) \end{bmatrix}. \quad (7)$$

After FDE, $R(0)$ is discarded because DC sub-carrier is not used and is made null at the transmitter in the proposed NDCS-DFTs-OFDM.

Then, the upper part signals, $R(1)$ to $R(M/2)$ and the lower part signals, $R(N-M/2)$ to $R(N-1)$ are put into the M point IDFT processor to convert them to the time domain signals, $r(n)$ (n=0~M-1), which is given by:

$$\begin{bmatrix} r(0) \\ \vdots \\ r(M-1) \end{bmatrix} = \frac{1}{M} \boldsymbol{D}_M{}^H \begin{bmatrix} R(1) \\ \vdots \\ R(M/2) \\ R(N-M/2) \\ \vdots \\ R(N-1) \end{bmatrix}, \quad (8)$$

where $\boldsymbol{D}_M{}^H$ is the complex conjugate of the transpose matrix of $\boldsymbol{D}_M$. The reconstructed signal, $r(n)$ is never affected by DC offset error between AD converters and I/Q demodulator, like OFDM signals of IEEE802.11a, for example.

C. Application of DFT spreading to OFDM-based WLAN

As shown in Fig. 1 (a), all of the functional blocks except M point DFT processor in the NDCS-DFTs-OFDM transmitter are included in IEEE802.11n/ac OFDM transmitter. Furthermore, as shown in Fig. 1(b), all of the functional blocks except M point IDFT processor in the NDCS-DFTs-OFDM receiver are also included in IEEE802.11n/ac OFDM receiver. Furthermore, all of the sub-carriers used in the proposed NDCS-DFTs-OFDM are exactly the same as those of conventional IEEE802.11n/ac, thus the preamble format used in IEEE802.11n/ac WLAN can be re-used and do not need to be changed at all. In addition, we can expect the same robustness against DC offset error as IEEE802.11n/ac OFDM.

D. Additional signal processing complexity

Additional signal processing complexity needs to be evaluated when the proposed NDCS-DFTs-OFDM is applied to OFDM based WLAN. Signal processing complexity of NDCS-DFTs-OFDM is essentially the same as that of conventional DFTs-OFDM. If M is power of two, we can employ FFT algorithm to reduce the number of complex multiplication for DFT. However, when we assume the existing parameters of IEEE802.11n/ac WLAN, the number of sub-carriers, M is 56 and the FFT size, N is 64 for 20MHz band operation of IEEE802.11n/ac. Table I shows the comparison of signal processing complexity by the number of complex multiplications of NDCS-DFTs-OFDM and OFDM. FFT algorithm can reduce 64^2 complex multiplications to 192 for 64 points DFT. On the other hand, PFA (Prime Factor FFT algorithm) can be applied for 56 points DFT to reduce the signal processing complexity [7]. PFA can reduce 56^2 complex multiplications to 476 for 56 points DFT. Therefore, the total number of complex multiplications required for NDCS-DFTs-OFDM is 668 at the transmitter as shown in Table I. The signal processing complexity of NDCS-DFTs-OFDM is about 3.5 times larger

 ISBN: 978-1-61208-279-0

than that of OFDM. This is a trade-off issue between the additional power consumption for DFT spreading in NDCS-DFTs-OFDM and the power consumption due to large OBO of HPA in OFDM. Considering that Moore's law is still effective, NDCS-DFTs-OFDM seems promising to improve energy efficiency of OFDM based WLAN in near future.

In IEEE802.11ah using UHF band, 1/10 clock down operation of IEEE802.11ac OFDM is proposed. In this case, power consumption of OFDM modem is essentially 1/10 compared with that of IEEE802.11n/ac and additional signal processing complexity will not be a significant problem. Therefore, NDCS-DFTs-OFDM will be useful for energy efficiency improvement.

III. PERFORMANCE EVALUATION

A. Simulation parameters

PAPR, power spectrum, BER performance and EVM of NDCS-DFTs-OFDM are evaluated by computer simulation and are compared with those of OFDM and DFTs-OFDM. Major simulation parameters are shown in Table II, where the number of sub-carriers and FFT/IFFT size are the same as those of IEEE802.11n/ac standard. BER performance is evaluated in AWGN (Additive White Gaussian Noise) channel to demonstrate its robustness to DC offset error.

B. PAPR

Fig. 2 (a) shows CCDF (Complementary Cumulative Distribution Function) of PAPR of the proposed NDCS-DFTs-OFDM and compares it with those of the conventional OFDM and DFTs-OFDM in the case of QPSK. As DFTs-OFDM is essentially the same as single carrier modulation signals filtered by the ideal filter, its PAPR is the least among three schemes, and is 2dB lower than that of OFDM at CCDF=1%. PAPR of the proposed NDCS-DFTs-OFDM is slightly higher than that of DFTs-OFDM, however it is 1.7dB lower than that of OFDM. Fig. 2 (b) compares CCDF of PAPR of the proposed NDCS-DFTs-OFDM with DFTs-OFDM and OFDM in the case of 16QAM. PAPR of NDCS-DFTs-OFDM is slightly higher that that of DFTs-OFDM, however it is 1.2dB lower that that of OFDM. These results confirm that the proposed NDCS-DFTs-OFDM scheme achieves as low PAPR as DFTs-OFDM.

C. Power spectrum

Fig. 3 shows power spectrum of NDCS-DFTs-OFDM in the case of QPSK and compares it with DFTs-OFDM and OFDM in the case of linear amplifier and non-linear amplifier with OBO=3dB. Rapp model is used for the simulation of non-linear amplifier. As seen in Fig. 3 (a), there is no difference among three schemes in a linear channel. Note that DFTs-OFDM has slightly narrower spectrum as DC sub-carrier is used. In the case of non-linear amplifier with OBO=3dB as shown in Fig. 3 (b), the side-lobe level of DFTs-OFDM signals is the least among three schemes. The side-lobe level of NDCS-DFTs-OFDM is almost the same as that of DFTs-OFDM and a few dB lower than that of OFDM. This means the proposed scheme has lower adjacent channel power leakage than OFDM when OBO is 3dB and can use smaller OBO than OFDM.

D. Power spectrum

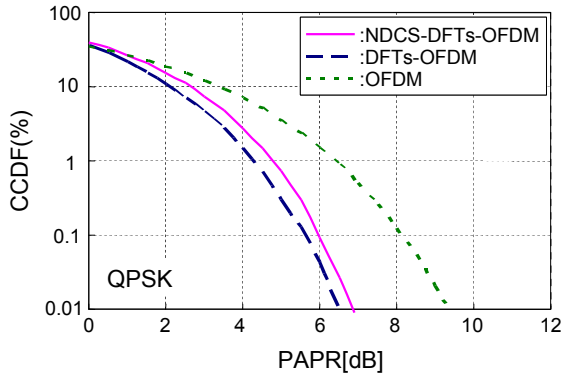
Fig. 3 shows power spectrum of NDCS-DFTs-OFDM in the case of QPSK and compares it with DFTs-OFDM and

TABLE I. SIGNAL PROCESSING COMPLEXITY IN TERMS OF THE NUMBER OF COMPLEX MULTIPLICATIONS AT THE TRANSMITTER

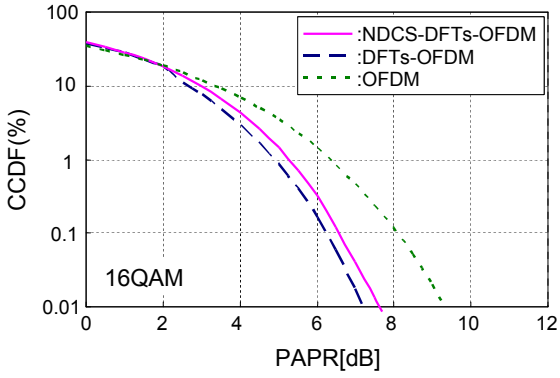
	DFT spreading (56 point)	IFFT (64 point)	Total
OFDM	0	192	192
NDCS-DFTs-OFDM DFTs-OFDM	476	192	668

TABLE II. MAJOR SIMULATION PARAMETERS.

Number of sub-carriers	56
FFT/IFFT size	64
Modulation scheme	QPSK/16QAM/64QAM
FEC	Convolutional-coding−Viterbi-decoding, R=3/4 or R=1/2



(a) QPSK



(b) 16QAM

Figure 2. PAPR comparison of NDCS-DFTs-OFDM, DFTs-OFDM and OFDM.

 ISBN: 978-1-61208-279-0

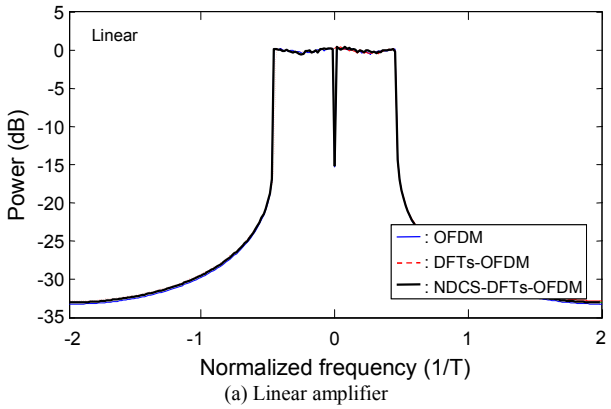
OFDM in the case of linear amplifier and non-linear amplifier with OBO=3dB. Rapp model is used for the simulation of non-linear amplifier [8]. As seen in Fig. 3 (a), there is no difference among three schemes in a linear channel. Note that DFTs-OFDM has slightly narrower spectrum as DC sub-carrier is used. In the case of non-linear amplifier with OBO=3dB as shown in Fig. 3 (b), the side-lobe level of DFTs-OFDM signals is the least among three schemes. The side-lobe level of NDCS-DFTs-OFDM is almost the same as that of DFTs-OFDM and a few dB lower than that of OFDM. This means the proposed scheme has lower adjacent channel power leakage than OFDM when OBO is 3dB and can use smaller OBO than OFDM.

E. BER performance and EVM

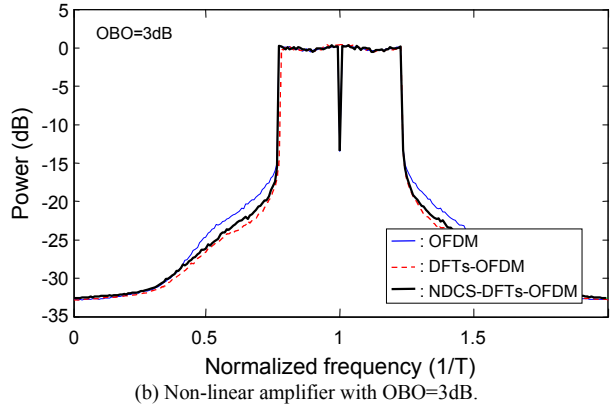
Fig. 4 compares BER performance of NDCS-DFTs-OFDM and DFTs-OFDM using QPSK with/without R=3/4 FEC, when DC offset error is 5%. DC offset error is defined by ΔA/A at the D/A converter of the transmitter and the A/D converter of the receiver where ΔA is DC offset error and A is the eye aperture at I/Q channels. When DC offset error exists, NDCS-DFTs-OFDM shows no E_b/N_o degradation and less E_b/N_o degradation at BER=10^{-4} in comparison with DFTs-OFDM.

Fig. 5 compares E_b/N_o degradation of NDCS-DFTs-OFDM and DFTs-OFDM as a function of DC offset error at BER=10^{-4}, where QPSK and 16QAM with/without R=1/2 FEC are employed for performance evaluation. Large E_b/N_o degradation due to DC offset error is observed in DFTs-OFDM. Though E_b/N_o degradation at BER=10^{-4} is as large as 3dB for 16QAM without FEC and 1dB for 16QAM with FEC in DFTs-OFDM, NDCS-DFTs-OFDM shows no E_b/N_o degradation due to DC offset error.

Another measure to evaluate the robustness against DC offset error is EVM. Fig. 6 compares EVM of QPSK,



(a) Linear amplifier



(b) Non-linear amplifier with OBO=3dB.

Figure 3. Power spectrum comparison of NDCS-DFTs-OFDM, DFTs-OFDM and OFDM

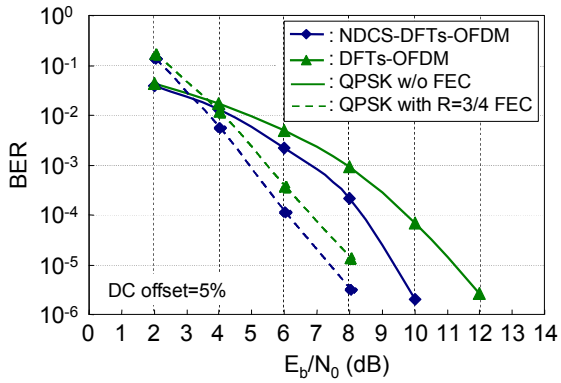


Figure 4. BER performance of NDCS-DFTs-OFDM and DFTs-OFDM when DC offset error=5%.

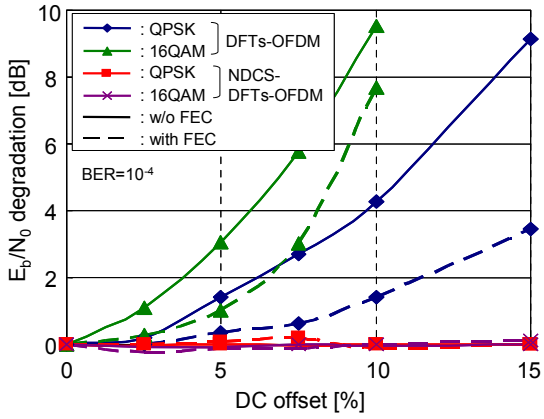


Figure 5. E_b/N_o degradation as a function of DC offset error in NDCS-DFTs-OFDM and DFTs-OFDM.

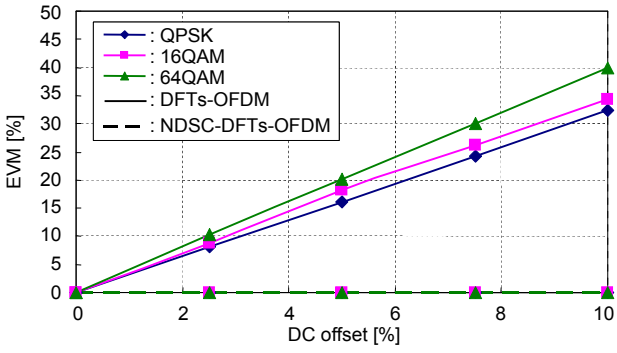


Figure 6. EVM as a function of DC offset error in NDCS-DFTs-OFDM and DFTs-OFDM.

 ISBN: 978-1-61208-279-0

16QAM and 64QAM as a function of DC offset error for NDCS-DFTs-OFDM and DFTs-OFDM. As shown there, EVM increases in proportion to DC offset error in DFTs-OFDM. EVM of 16QAM and 64QAM is slightly worse than that of QPSK. On the other hand, no EVM increase is shown according to DC offset error in NDCS-DFTs-OFDM, Therefore, BER performance of NDCS-DFTs-OFDM is not degraded due to DC offset error as shown in Fig. 4 and Fig. 5.

IV. CONCLUSIONS

This paper described an application of DFT spreading to OFDM based WLAN to improve energy efficiency. In order to maintain the robustness to DC offset error, this paper proposed a new type of DFTs-OFDM, NDCS-DFTs-OFDM by splitting the spectrum after DFT spreading into two parts and making DC sub-carrier null. Though the signal processing complexity is 3.5 times larger than OFDM under the condition that the basic parameters of OFDM are maintained, the simulation results confirmed the proposed NDCS-DFTs-OFDM achieves both advantages of low PAPR and robustness to DC offset error. Future work includes BER performance evaluation of NDCS-DFTs-OFDM in frequency selective fading as well as under the non-linear amplifier operation. In addition, the power consumption trade-off between NDCS-DFTs-OFDM and larger OBO of HPA must be conducted.

ACKNOWLEDGEMENT

A part of this work is supported by the Ministry of Internal Affairs and Communications on research and development for radio resource enhancement.

REFERENCES

[1] IEEE P802.11ad/D9.0, Draft standard, Part 11: Wireless LAN MAC and PHY Specifications, Amendment 4: Enhancements for Very High Throughput in the 60GHz Band.

[2] IEEE P802.11ac/D4.0, Draft standard, Part 11: Wireless LAN MAC and PHY Specifications, Amendment 4: Enhancements for Very High Throughput for Operation in Bands below 6 GHz.

[3] M. Park, “Specification Framework for TGah,” IEEE802.11TGah document, 11-11-1137-10-00ah, 2012.

[4] B.E. Priyanto, H. Codina, S. Rene, T.B. Sorensen, and P. Mogensen, "Initial Performance Evaluation of DFT-Spread OFDM Based SC-FDMA for UTRA LTE Uplink," IEEE Vehicular Technology Conference 2007 (VTC2007-Spring), 22-25 April 2007, pp. 3175-3179.

[5] H.G. Myung, J. Lim, and D.J. Goodman, "Peak-To-Average Power Ratio of Single Carrier FDMA Signals with Pulse Shaping," IEEE International Symposium on Personal, Indoor and Mobile Radio Communications 2006 (PIMRC 2006), 11-14 Sept. 2006, pp. 1-5.

[6] H.G. Myung, J. Lim, and D.J. Goodman, "Single carrier FDMA for uplink wireless transmission," IEEE Vehicular Technology Magazine, Vol.1, Issue 3, Sept. 2006, pp. 30-38.

[7] P. Duhamel and M. Vetterli, "Fast Fourier Transforms: A tutorial on fast Fourier transforms," Elsevier Science Publishers, Signal Processing, Vol. 19, Issue 4, April 1990, pp. 259–299.

[8] C. Rapp, “Effects of HPA-Nonlinearity on a 4-DPSK/OFDM-Signal for a Digitial Sound Broadcasting System”, Proceedings of the Second European Conference on Satellite Communications, Liege, Belgium, Oct. 22-24, 1991, pp. 179-184.

 ISBN: 978-1-61208-279-0

Static Bluetooth Scatternet Formation Models: The Impact of FHSS

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 potential of Wireless Personal Network (WPAN) applications is virtually untapped. The presence of low cost Bluetooth technology in most mobile devices makes it the logical choice for further exploration. Existing Bluetooth models neglect its own communication technique, the Frequency-Hopping Spread Spectrum (FHSS), and part of the challenge of generating efficient Scatternet algorithms is related to the characteristics of FHSS. We propose a model of FHSS using dynamic graphs and show the impact of its use in topology of these networks. We show that small changes in a centralized model of Scatternets lead to better traffic and consumption; also, their topology results are comparable to those of dynamic models.**

***Keywords*—bluetooth; scatternet; fhss; dynamic graph.**

I. INTRODUCTION

Applications using Wireless Personal Area Network (WPAN) have not yet explored their full potential. The possibility of forming wider-ranging Ad Hoc networks among low cost and low power consumption devices enhances their most common use, which is limited by the data cable replacement. Some of the possible applications of WPAN are: profile, message and location exchange in mobile social networks; mesh networks for Internet access using mobiles as repeaters; networks for monitoring life support medical devices; residential (smart buildings) and industrial automation; dating services and social games, by connecting users with compatible profiles using an open network of mobile devices.

Bluetooth is the most popular WPAN technology [1]: 906 million mobile phones were sold in 2010, almost all with Bluetooth; 171 million laptops were shipped in 2010, 77% of them with Bluetooth; more than 50 million game consoles were shipped in 2010, 62% of them with Bluetooth; more than 40 million Bluetooth enabled health and medical devices were sold in early 2011; one third of all new vehicles produced in 2011 included Bluetooth, which will increase to 70% by 2016.

Studies carried out between 2002 and 2006 [2] introduced different Bluetooth Scatternet formation protocols. Due to the increasing use of Bluetooth, current research has once again focused on the Bluetooth Scatternet formation protocols: [3]–[5], to list a few. However, none of them have been standardized yet and, therefore, no commercial products include this functionality. In the context of mobile social networking applications for example, we found no popular application despite the increasing number of devices manufactured with Bluetooth. This void is due to the complexity of implementing Scatternet algorithms.

Bluetooth allows devices to communicate using channels with a hopping sequence coordinated by the master device and known to all participants in the Piconet. In order to establish a connection, the pattern of frequency hops must be known, a technique called Frequency Hopping Spread Spectrum (FHSS).

The Bluetooth connection process involves two phases: discovery and link formation. During the discovery phase, the device that will assume the role of master scans for slave devices waiting for connection. Master and slaves begin a sequence of pseudo-random frequency hops, until a frequency coincidence occurs. After the match takes place, the link formation phase starts with the slave waiting for a random time to respond to the master: the Backoff interval. Randomness and a Backoff time are necessary to avoid collisions, however, they introduce seconds of delay to the initial connection.

On evaluation of the efficiency of the Scatternet, the FHSS is seen to exert a significant influence on the Bridge nodes, which are responsible for the inter-Piconet communication. In order for a node to act as a Bridge, it must stop communicating within one Piconet and change its standard hopping sequence for another.

During this procedure, the node that plays the role of the Bridge enters into the HOLD state. This procedure has a high associated cost due to the master exchange and subsequently the exhange of hopping sequence and the synchronisation of responses, which are coordinated by the Scheduling. These characteristics mean that the location, volume and types of these Bridges directly influence energy consumption and traffic.

Assessment and understanding of the Scatternet algorithms and models are essential, because many of the delays and losses in performance and energy observed during the formation and coordination were attributed to the complexity of the algorithm that can generate many Bridges, and the need for the discovery of new nodes.

Our study focuses on two issues related to the use of the FHSS in Bluetooth: the randomness of the discovery procedure and the influence of Bridges on the performance of a Scatternet.

We show that there is a substantial degradation of the de-

 ISBN: 978-1-61208-279-0

vice discovery procedure, and we analyze how this degradation affects the classic static Bluetooth Scatternet formation. This problem is directly related to the FHSS. According to Jedda et al. [7], [8], this is the main cause of the lack of efficiency in the Bluetooth Scatternet Formation protocols and the absence of its adoption in the standardization process.

Our contribution to this field of research:

- **Model Bluetooth as a Dynamic Graph for use in Static Bluetooth Scatternet Formation Model**: From the work of Pettarin et al. [9], which models Bluetooth as a graph, we improved this model with the *FHS() process* - representing the Frequence Hopping Sequence, and the *Disc()* - representing the Discovery process. These two new functions characterize the Bluetooth graph as a dynamic graph. This model can be used as a requisite to create static Bluetooth Scatternet formation models;
- **A New Optimization Model for Static Scatternet Formation**: We propose an update of a classic static Bluetooth Scatternet formation model, the mathematical programming model described in Marsan et al. [6]. We create a new model by penalizing the activation of Bridges and by including new constraints. This new model produced optimal solutions in which the structure of the Scatternets is more coherent with the ones predicted by Law et al. [10], a well-known dynamic Bluetooth Scatternet Formation model. Our considerations can be used as constraints for other static Bluetooth Scatternet formation models;

Section II shows the related work; Section III-A explores WPANs, detailing the machine in the connection state of Bluetooth. In Section III-B, we model Bluetooth FHSS in a dynamic graph. In Section III-C, we introduce the Scatternet, and in Section III-D, we address the elements that influence the efficiency of its formation. In Section IV, we analyze the effects on FHSS in our experiments. Finally, we detail the conclusions in Section V.

II. Related Work

In Tahir et al. [5] the Scatternet routing protocol (SROP) is proposed. Its main purpose is to establish routes with a minimum number of hops between source and destination. For this to occur, the master searches all possible paths through flooding. Its main advantage is the fact that since the master knows all the links, it can use this information if any device at the source wants to communicate with a member of the Piconet. According to the author, there is a time saving benefit in this search for routes using flooding, and its performance is evaluated in the ns-2 [11] simulator.

Jedda et al. [7] analyzed the impact of Bluetooth specification parameters on the convergence of a Scatternet. These impacts are related to the use of the FHSS communication technique. Using the ns-2 [11] simulator, the differences in convergence times between dynamic and static algorithms in Bluetooth Scatternet Formation (BSF) are shown. Some results showed that changes in the implementation of Bluetooth are more significant in static algorithms.

Pettarin et al. [9] discuss the expansion and diameter of the ad-hoc Bluetooth topology induced by the discovery phase, by means of a Random Geometric Graph (RGG). However, the work does not explore the topology of a Piconet or the intrinsic characteristics of FHSS.

In Law et al. [10] a new dynamic algorithm of Scatternet formation is introduced. In its organization, devices are separated by components. It defines a device, Piconet or Scatternet as a component. They show that their algorithm has $O(\log n)$ time complexity and $O(n)$ message complexity, which generates an algorithm with an efficient battery usage.

In Chiasserini and Marsan [12] the restrictions of the centralized model of Marsan et al. [6] are complemented by a discussion and proposal of distributed algorithms in Scatternet formation, including routines for the insertion and removal of nodes.

Marsan et al. [6] provide a description of the Scatternet formation using mathematical programming. Constraints are proposed in a min-max formulation, leading to an optimization problem, which is solved in a centralized way. However, this article does not take into account the effects of FHSS. This approach is limited to the rules established for each device, without assigning penalties for Bridge nodes.

III. Methodology

A. WPANs

Wireless Personal Area Networks (WPANs) are wireless networks between low cost, energy-consumption and data-loss devices that create short links around the user's workspace. Bluetooth is the most common WPANs technology, whose communication uses FHSS and has a connection range of 10 m, as found in the most common versions present in the market. FHSS is a common communication technique in ad-hoc peer-to-peer networks. They communicate on one channel for each time slot. They are less susceptible to noise from neighboring networks, can be used at various distances, offer QOS and stronger security compared to traditional 802.11 Wi-Fi. A Bluetooth network is called a Piconet, and its nodes act as master or slaves.

Two distinct phases are required to connect Bluetooth devices: the Discovery and the Link formation. During the Discovery phase, the device that will assume the role of master goes into the INQUIRY state, looking for slave devices awaiting a connection in an INQUIRY SCAN state.

The searching device sends an identifier called Inquiry Access Code (IAC). During the INQUIRY, the IAC is broadcast on 32 of the 79 frequencies defined by the specification, divided into two trains of 16. This sequence of hopping frequencies occurs in a pseudo-random way, with calculations derived from the clock of the device. Master and slaves begin a sequence of pseudo-random frequency hops, until a frequency coincidence occurs.

A time slot difference collaborates with the increased likelihood of the device hearing the same channel on which a IAC was transmitted: the devices in INQUIRY state hop in time slots of $312.5\mu S$ faster than the standard Bluetooth $625\mu S$ used by devices in INQUIRY SCAN.

 ISBN: 978-1-61208-279-0

After receiving an IAC, the slave device assumes a state called INQUIRY RESPONSE, waits for the Backoff time to elapse and responds to the request by sending its network address and clock in a packet called Frequency Hopping Synchronization (FHS). After this process, it enters a state called PAGE SCAN. When the master receives the FHS, it enters a state of PAGE, and uses the information received from the FHS for synchronization and connection with the slave nodes that have already been discovered and are in the PAGE SCAN state.

During Backoff, the device waits for a random value of Time slots $(0 - 639.375)\mu S$. This status is set after the device receives a master's IAC, with the objective of minimizing packet collisions of response to the master. When the Backoff time is over, the device waits for a new IAC in the INQUIRY RESPONSE state, by sending its hop pattern and clock to the master.

During the PAGE state, the master device selects a slave to be connected through its network address, and sends packages through the sequence of estimated hops in the clock of the slave previously discovered.

After the PAGE process is complete, the Piconet is formed and the devices gain an online status and may also negociate the roles of master and slave.

Intra-Piconet communication requires a Scheduling process, during which the master performs a polling on each slave, and only upon receipt of this packet shall they be allowed to communicate again in the Piconet. The order in which the slaves receive this package is called Polling Cycle and it also determines the slots that will be used. This scheduling must be coordinated by an algorithm that determines the sequence in which the master will poll the slaves.

B. Dynamic Graphs

The Bluetooth network will be described as a graph, according to the definition of Gupta and Kumar [13]. According to Pettarin et al. [9], the links can be described by the function $c(n)$, where n is the number of devices, and the range of each device is $r(n)$ With this, the Bluetooth network denoted by the graph $BT(r(n), c(n))$, where $r(n)$ are the vertices V_n and $c(n)$ are the edges E_n. A set of nodes, V_n , with spatial displacement described by a random variable N , uniformly distributed in $[0, 1]^2$; a set of edges E_n , obtained as follows: each vertex, $u \in V_n$, selects a random set of $c(n)$ neighbours, all at the distance of $r(n)$; one edge $e_i = \{u, v\}$, $e_i \in E_n$ exists only if the vertex u selects another vertex v.

Pettarin et al. [9] describes a situation in which the devices are moving and setting the Bluetooth network connection, as a sequence of graphs $G_t(n, \rho, r(n), c(n), t)$, in which ρ is the set of nodes that are part of the connection graph, and t is each time step, linked to the movement of devices. This sequence of graphs can be compared to a Markov chain, whose transition can be described by the model of moving nodes or by spontaneous disconnections.

Assuming that F is the set of frequencies used in FHSS, so that $f_i \in F, 0 < i \leq 79$. FHS is a function of $FHS(CLK, MS)$, where CLK is the clock of the elements involved and MS is the address of the Piconet master. The details of FHS are given by Bluetooth specifications. We have F', so that $f_i = FHS(CLK, MS)$, which is an pseudo-random sequence of the set F. The sequence of F' is unique to each master-slave link. Thus, for communication to take place, each $e_i = \{u, v\}$ must have the same F'.

Let u and v be vertexes that meet the formalization described in the previous section. We define the discovery process as the operation $Disc(u, v, f_i)$, an operation to insert an edge from the set E_n. The $Disc()$ process has its execution distributed, while running u and v at the same time. Master and slaves begin a sequence of pseudo-random frequency hops, until a frequency f_i coincidence occurs. Once matched, the slave waits for a random time to respond $FHS()$ to the master, and this is called the Backoff interval. This is necessary because the $FHS()$ must be exchanged between nodes. For this reason, after $Disc()$ the slave returns $FHS()$ and generates the correct F' for the connection. MS will belong to the element that has been selected as the master.

Given this definition, the graph $BT(r(n), c(n))$ can be classified as a dynamic graph. According to Frigioni et al. [14], a dynamic graph is a graph G whose edges are not fixed and in which some property p of a given graph $G = (V_n, E_n)$ is considered to be true after a series of operations. The algorithm that maintains this property classifies the vertices and edges in different states, with operations that alter these states. These sequences and operations define a dynamic model of edges, which can be removed and inserted into E_n during the verification of F'.

C. Scatternets

Scatternets are collections of Piconets that are formed spontaneously without fixed infrastructure. They are dynamic networks that enable nodes to communicate in scenarios of more than one hop. They break the centralized limits of the *Bluetooth* specification with star topology, coordinated by a master, thus making mesh formations possible. The Scatternet formation rules do not receive further details from the Bluetooth specification, thus enabling other alternatives to be created by means of distributed algorithms, which establish the rules of Piconet associations.

The elements that enable multihop communication across the Scatternet are called "Bridges". 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.

The HOLD mode permits FHSS nodes to be Bridges and to interact with other networks. During the HOLD mode, an exchange of the node's master takes place and hence, a synchronization of the channel with the Piconet must occur. This way, a node leaves its Piconet, changes its hopping pattern, and starts to receive Pollings from the master of another Piconet. It should be noted that a device cannot be the master of more than one network, but may be master in one Piconet and a slave in multiple Piconets, acting as a Bridge. This procedure has a cost associated to the change of hopping pattern from another Piconet and with Scheduling. These costs

 ISBN: 978-1-61208-279-0

directly influence the optimization of Scatternet, which in turn, directly influences energy consumption and traffic.

Now we introduce the static and dynamic models.

1) Static Model: The centralized 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 optimization problem which is solved in a centralized way. It can find the best possible performance for a given graph, obeying the Piconet Bluetooth restrictions. The objective of this model is to minimize 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.

Marsan et al. [6] model, for instance, discuss the centralized 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.

These requirements and constraints lead to a min-max criterion which is solved using CPLEX [15]. In its model of constraints, the pseudo-randomness of the Discovery phase is not included. The influence of the delays caused by the effort involved in switching channels during traffic between Piconets by a Bridge node in HOLD mode is not addressed either.

2) Dynamic Model: Dynamic models of Scatternets are protocols, and its distributed algorithms use the following heuristic [2]:

- 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 minimizes 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.

Law et al. [10] show in their dynamic model that their algorithm has $O(\log n)$ time complexity and $O(n)$ message complexity. However, according to Jedda [8], their dynamic model also does not consider the improvements made to FHSS, which appear in Bluetooth version 1.2.

D. Efficiency In Scatternet Formation

The location of Bridges is critical for the evaluation of the impact of the resulting topology. Given they are responsible for the inter-Piconet communication, they are subjected to more communication overhead and processing than other nodes.

We evaluated the influence of the number of Bridges on a Scatternet and the need of the HOLD mode for data to be exchanged between Piconets. The HOLD mode consumes energy and influences traffic, being one of the states that require the greatest effort in the resynchronization of frequencies in the Piconets, and which participates in and awaits communication polling during the Scheduling process. An efficient Scatternet topology should minimize its use because:

1) Fewer Bridges mean less delays in migration to another Piconet during the transmission of the received data;
2) Fewer masters on the network imply fewer Bridges, and fewer delays in the exchange and resynchronization of the channels of the new Piconet.

IV. RESULTS

We used the UCBT [16], an extension that simulates Bluetooth in ns-2 [11], developed by the University of Cincinnati.

In order to assess the delay in the formation of new Piconets during the Discovery phase, we generated instances with 1 master and $c(n)$ neighbouring devices, which were candidates for the role of slave.

In accordance with Pettarin et al. [9], we observed that as the value of $c(n)$ is increased, so is the likelihood of connection. The rationale for this is that during the Discovery phase, all devices in INQUIRY SCAN perform pseudo-random hops in slower Time Slots than the master until there is a match of frequencies. This behavior shows that, despite the increase in density of devices within the range of the master, the FHSS provides greater resilience to collisions and depletion of the spectrum.

Figure 1 shows the formation of a theoretically maximal Piconet, represented by one master and seven slaves. We observed the proportional increase value of $c(n)$ and the time, until it forms a Piconet with 1 master and 7 slaves. This behaviour is explained by the need for matching the channel in the Discovery phase, the Backoff and Scheduling of intra-Piconet packets.

In order for a new discovery to take place, the master needs to stop the intra-Piconet comunication. While the slaves that have already entered the Piconet change to the HOLD mode, waiting for new pollings from the master before recommunicating. The time cost of this operation grows with the

 ISBN: 978-1-61208-279-0

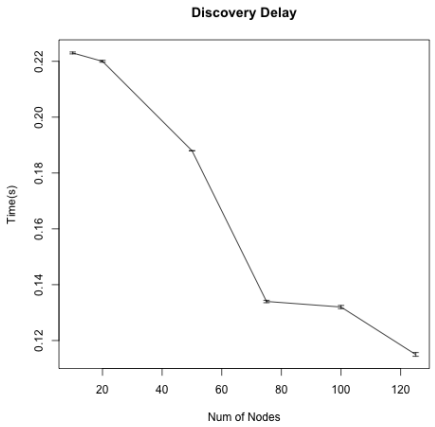


Fig. 1. Time of first master-slave connection

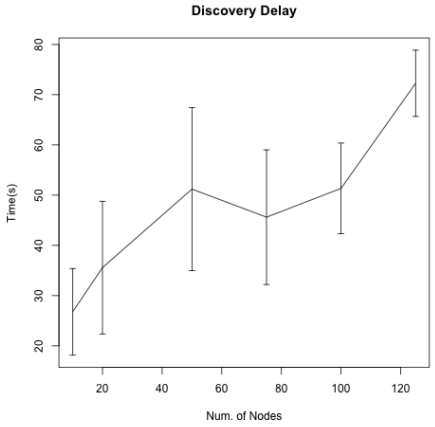


Fig. 2. Time until formation of the first complete Piconet with 7 slaves and 1 master

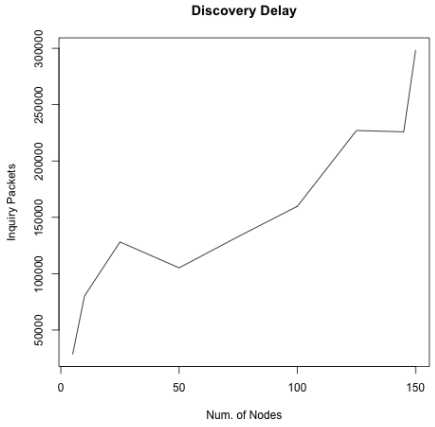


Fig. 3. INQUIRY packets necessary for formation of the first Piconet

increase of devices $c(n)$ due to the randomness of the intra-Piconet Scheduling and Discovery of new slaves. The error bars in Figure 2 show the high degree of variability and delay in connection, represented by random variables associated with the Discovery of slaves, Backoff time and Intra-Piconet Scheduling processes.

Figure 3 shows the number of INQUIRY packets transmitted as a function of the number of devices in the formation of the first Piconet.

A. *Topology analysis*

To evaluate the efficiency of a Scatternet topology and the influence of Bridges, we simulated the formations with Law et al. [10] algorithms. We generated 30 instances of 20 devices until full convergence. Based on the results, we generated graphs of the most common Scatternet obtained with 20 nodes, Figure 4.

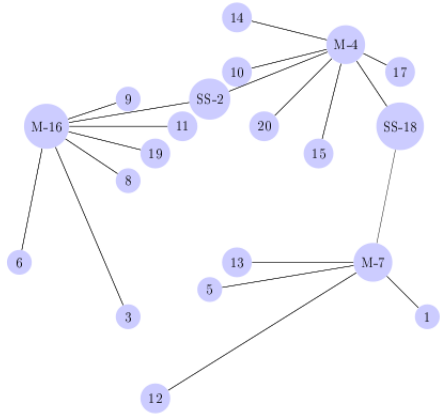


Fig. 4. Common Scatternet of 20 devices found by Law et al. [10] algorithm

The graphs generated in our model follow the rules of efficiency as per the dynamic algorithm of Law et al. [10]; centralized models like Marsan et al. [6] need to be modified so that the results are closer to an efficient energy consumption result.

Looking at the graph shown in the article Marsan et al. [6] Figure 5, we found 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;
- Node 9 is the Bridge of three Piconets, a prohibitive result;
- We observe various network loops between the Piconets of masters 7 and 17, connected by nodes 9 and 15;
- Four Piconets is an excessive amount for 20 nodes.

B. *Improving the centralized model of Marsan et al. [6]*

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;

 ISBN: 978-1-61208-279-0

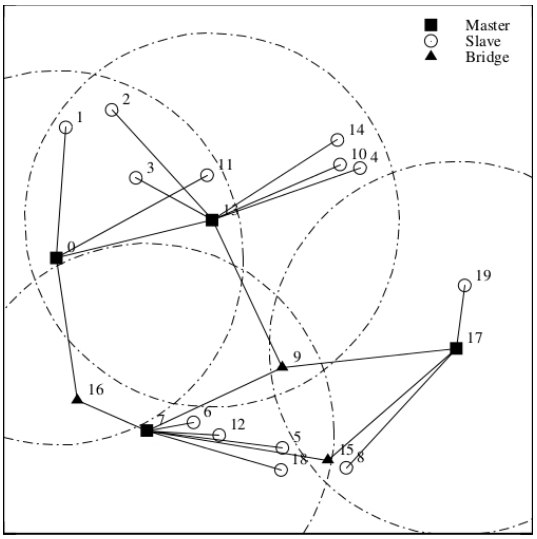


Fig. 5. Scatternet with 20 devices found in Marsan et al. [6] model

- 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, which are equal to 1 if the node is a master, a Bridge or a slave, respectively, and are otherwise equal to 0. 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 I :

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

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

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

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

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

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

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

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

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

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

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

In order that the solution to the problem of the centralized model of Marsa et al. [6] generates a topology similar to that obtained by simulation using the dynamic algorithm of Law et al. [10], we had to add two new constraints to the eleven existing in the original model.

TABLE I. 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.

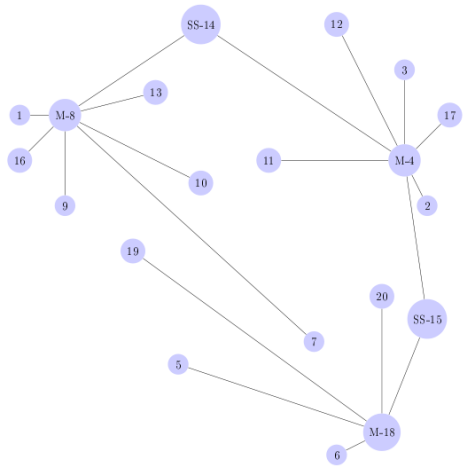


Fig. 6. Scatternet generated with modified model of Marsan et al. [6]

- $\mu_i + \mu_j + x_{i,j} \leq 2 \quad \forall\, i,j \in N \quad i \neq j$ **; 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 to the Bridges and these two constraints, we make sure that the resulting graph has a topology that is less prone to effects resulting from delays when Bridges are in HOLD mode. Considering a dynamic graph, these penalties are associated to the cost of functions *Disc()* and *FHS()* shown in Section III-B.

To demonstrate the efficacy of our modification, we used the same instance as that of Marsan et al. [6], with input parameters equal to those of Table II, which form the graph of Figure 5.

TABLE II. INPUT PARAMETERS

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

We can see in the Scatternet topology found by our model in Figure 6, that all the items required for effective formation, as previously mentioned in Section III-D, are respected. Our static model approaches the results of Law et al. [10] as

 ISBN: 978-1-61208-279-0

shown in Figure 4. This algorithm has a cost of $O(\log n)$ time complexity and $O(n)$ message complexity, so we can conclude that our resulting graph represents a Scatternet with an ideal distribution of data flow and power consumption.

V. DISCUSSION

The delay of the Bluetooth connection process and the loss of efficiency of some of the algorithms in Scatternet formation are directly related to the effects of FHSS. A correct mapping of its peculiarities is essential for the suitable design of Bluetooth solutions and applications.

The delay in Discovery is the determining factor for simpler applications restricted to a Piconet that requires adequate responsiveness. The *Disc()* and *FHS()* functions in the dynamic graph, shown in Section III-B, models the demand of this factor and its importance in the Scatternet formation. We simulated the formation of Piconets and Bluetooth Scatternets to analyze the delay caused by the discovery process for new nodes during the formation of a new Piconet, and the entry of new slaves into an existing Piconet.

The centralized 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 in the constraints, we achieved results similar to those obtained by simulation of another classic 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. [10]: complexity of $O(\log n)$ time complexity and $O(n)$ message complexity. Our optimization can be used as a requisite for other static Bluetooth Scatternet formation models.

Research that proposes changes to the Bluetooth specification or workarounds for some yet-to-be-explored use examples are a necessity, given the context of the increasing popularity of Bluetooth in Smartphones and Tablets, thereby leveraging this new wave of applications, which is still virtually unexplored due to the side effects of using FHSS.

In future work, we will state that both static and dynamic Bluetooth Scatternet Formation protocols must consider the impact of the FHSS for achieving more practical results and process standardization.

ACKNOWLEDGEMENT

Thanks to Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), to CNPq, Capes and FAPEAL.

REFERENCES

[1] Bluetooth.com, "The bluetooth network effect," Last Visited in 11/02/2013.[Online]. Available: http://www.bluetooth.com/Pages/network-effect.aspx/

[2] R. Whitaker, L. Hodge, and I. Chlamtac, "Bluetooth scatternet formation: A survey," 2004.

[3] C. M. Yu, "Gloal configured method for blueweb routing protocol," IET Communications, vol. 6, no. 1, january 2012, pp. 69–75.

[4] S. Tahir, A. M. Said, and S. T. Bakhsh, "Bluetooth network re-formation protocol for reducing path length (bnr)," in Computer Information Science (ICCIS), 2012 International Conference on, vol. 2, june 2012, pp. 755–759.

[5] S. Tahir, A. M. Said, and S. T. Bakhsh, "A Bluetooth Scatternet Route Optimization Protocol," in AASRI Procedia - AASRI Conference on Power and Energy Systems, vol. 2, 2012, pp. 142–148.

[6] M. A. Marsan, C. Chiasserini, A. Nucci, G. Carello, L. D. Giovanni, and L. D. Giovanni, "Optimizing the topology of bluetooth wireless personal area networks," 2002, pp. 572–579.

[7] 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.

[8] A. Jeddah, N. Zaguia, and G.-V. Jourdan, "Analyzing the device discovery phase of bluetooth scatternet formation algorithms," in Distributed Computing Systems Workshops, 2009. ICDCS Workshops 09. 29th IEEE International Conference on, june 2009, pp. 468–471.

[9] A. Pettarin, A. Pietracaprina, and G. Pucci, "On the expansion and diameter of bluetooth-like topologies," in Algorithms - ESA 2009, ser. Lecture Notes in Computer Science, A. Fiat and P. Sanders, Eds. Springer Berlin / Heidelberg, 2009, vol. 57, pp. 528–539.

[10] C. Law, A. K. Mehta, and K.-Y. Siu, "A new bluetooth scatternet formation protocol," Mob. Netw. Appl., vol. 8, no. 5, Oct. 2003, pp. 485–498,

[11] N. S. 2, "The network simulator - ns2," Last Visited in 11/02/2013. [Online]. Available: http://www.isi.edu/nsnam/ns/

[12] C. F. Chiasserini and M. A. Marsan, "Towards feasible topology formation algorithms for bluetooth-based wpans," in in Proc. Hawaii Int. Conf. on System Science, Big Island. Society Press, 2003, pp. 313–322.

[13] P. Gupta and P. Kumar, "The capacity of wireless networks," Information Theory, IEEE Transactions on, vol. 46, no. 2, , Mar. 2000, pp. 388–404.

[14] D. Frigioni and G. F. Italiano, "Dynamically switching vertices in planar graphs (extended abstract)," in Proceedings of the 5th Annual European Symposium on Algorithms. London, UK: Springer-Verlag, 1997, pp. 186–199.

[15] IBM, "Ibm ilog cplex optimizer," Last Visited in 11/02/2013. [Online]. Available: http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/

[16] D. A. Q. Wang, "Ucbt - bluetooth extension for ns2 at the university of cincinnati," Last Visited in 11/02/2013. [Online]. Available: www.cs.uc.edu/cdmc/ucbt/

 ISBN: 978-1-61208-279-0

Performance Analysis of Complex Combiner at Two Time Instants in Weibull Fading Channel

Dragana Krstić
Department of Telecommunications,
Faculty of Electronic Engineering,
University of Niš
Niš, Serbia
dragana.krstic@elfak.ni.ac.rs

Goran Stamenović
Tigar, Pirot, Serbia
goran.stamenovic@tigar.com

Petar Nikolić
Tigartyres,
Pirot, Serbia
nikpetar@gmail.com

Aleksandar Stevanović
Mechanical Technical School 15th of May
Niš, Serbia
aleksandar.stevanovic_mts@yahoo.com

Abstract **—The expressions for joint probability density function (PDF) of the Switch and Stay Combiner (SSC) output signal at two time instants in the presence of Weibull fading are determined in the closed form. Then, in this paper, these equations are used for calculation of the outage probability and amount of fading for complex Switch and Stay Combining/Maximal Ratio Combining (SSC/MRC) combiner versus different parameter values. The results are shown graphically in some figures and the analysis of the parameters influence and different types of combiners is given.**

Keywords - Probability Density Function; Joint Probability Density Function; Outage Probability, Amount of Fading, Weibull Fading; SSC/MRC Combiner

I. INTRODUCTION

In wireless communication, the main causes of signal degradations are random fluctuations of signal envelope and phase, which are caused by multipath scattering (fast fading) and shadowing (slow fading) [1]. The multipath fading is modeled by several distributions such as: Rayleigh, Rice, Nakagami-*m*, The Hoyt (Nakagami-*q*), Weibull.

The Weibull fading, named after Waloddi Weibull, is a simple statistical model of fading, is based on the Weibull distribution and used in wireless communications [2]-[4], particularly with mobile radio systems operating in the 800/900 MHz frequency range [5]. Empirical studies have shown it to be an effective model in both indoor [5] and outdoor [6] environments.

Theoretical model for a particular class of Weibull distributions was described by Sagias and Karagiannidis [7]. Also, they analyzed the channel capacity of wireless channels in the presence of Weibull fading [8].

Various techniques for reducing fading and shadow effects are used in wireless communication systems. They are diversity reception, dynamic channel allocation and power control. Upgrading transmission reliability and increasing channel capacity without increasing transmission power and bandwidth is the main goal of diversity techniques.

Multiple received copies of signal could be combined in different ways. Among the most popular are maximal ratio combining (MRC) and equal gain combining (EGC) [9]-[11]. Their complexity of implementation is relatively high since they require a separate channel for each diversity branch. Selection combining (SC) and switch and stay combining (SSC) are simpler diversity combining models, because they process only one from *L* diversity branches. SSC combiner usually chooses between two receiving antennas based on comparation of the signal value or SNR of connected antenna with previously determined threshold. This is reflected in less complexity toward SC combiner and it is not necessarily at the same time continued to monitor the two antennas. Because of that there is some loss in performances.

The probability density function (PDF) of the SSC combiner output signal at one time instant and the joint probability density function of the SSC combiner output signal at two time instants in the presence of Weibull fading are determined in [12].

The authors showed that the error probability and the outage probability are significantly reduced if the decision is performed in two time instants. The analysis of the complex SSC/SC combiner over outage probability at two time instants in the presence of Rayleigh and log-normal fading is done in [13] [14] and the bit error rate for complex SSC/MRC combiner at two time instants in the presence of Rayleigh, Nakagami-*m*, Hoyt and log-normal fading is done in [15]-[18], respectively.

The bit error rate for complex SSC/MRC combiner at two time instants in the presence of Weibull fading will be given in this paper since Weibull fading has great importance in the study of telecommunications systems. This investigation could be useful for designers of wireless telecommunication systems.

 ISBN: 978-1-61208-279-0

This paper is organized as follows: after Introduction, Section II introduces the model of complex combiner which performances will be considered in the next section. In Section III, the joint PDF of the SSC combiner output SNR at two time instants is calculated. Subsequently, in fourth Section the outage probability and amount of fading for complex Switch and Stay Combining/Maximal Ratio Combining (SSC/MRC) combiner are calculated and then in the fifth Section the numerical results are presented graphically. Final part of this paper is conclusion with an analysis of the obtained results.

II. SYSTEM MODEL

The complex SSC/MRC combiner, considered in this paper, is presented in Fig. 1. The complex combiner is with two inputs, at two time instants. The SSC combiner output signals are the input signals for MRC combiner.

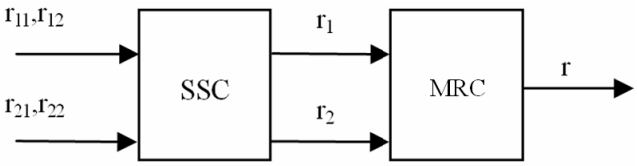


Figure 1. Model of complex combiner

At the inputs of the first part of complex combiner the signals are r_{11} and r_{21} at first time moment and they are r_{12} and r_{22} at second time moment. The first index represents the branch ordinal number and the other one signs the time instant observed.

The output signals from SSC part of complex combiner are r_1 and r_2. The indices at the output signal correspond to the time instants considered. These signals, r_1 and r_2, are the inputs for the MRC combiner. Finally, the overall output signal is r.

III. PERFORMANCE DERIVATION

The joint probability density function of correlated signals r_1 and r_2 at the SSC combiner output, at two time instants, Weibull distributed and with same parameters Ω_i and β_i [19], can be obtained in closed form from expressions in [12, eq. (15)-(18)] as

For $r_1<r_t$, $r_2<r_t$:

$$p_{r_1r_2}(r_1,r_2) = P_1 A(r_1,\beta_2,\Omega_2)A(r_2,\beta_1,\Omega_1) + P_2 A(r_1,\beta_1,\Omega_1)A(r_2,\beta_2,\Omega_2) \quad (1)$$

For $r_1\geq r_t$, $r_2<r_t$

$$p_{r_1r_2}(r_1,r_2) = P_1 A(r_1,\beta_1,\Omega_1)\frac{\beta_2}{\Omega_2} r_2^{\beta_2-1} e^{-\frac{r_2^{\beta_2}}{\Omega_2}} + P_1 A(r_1,\beta_2,\Omega_2)A(r_2,\beta_1,\Omega_1) + P_2 A(r_1,\beta_2,\Omega_2)\frac{\beta_1}{\Omega_1} r_2^{\beta_1-1} e^{-\frac{r_2^{\beta_1}}{\Omega_1}} + P_2 A(r_1,\beta_1,\Omega_1)A(r_2,\beta_2,\Omega_2) \quad (2)$$

For $r_1<r_t$, $r_2\geq r_t$

$$p_{r_1r_2}(r_1,r_2) = P_1\left(1-e^{-\frac{r_t^{\beta_1}}{\Omega_1}}\right)\frac{\beta_2^{\,2}(r_1r_2)^{\beta_1-1}}{\Omega_2^{\,2}(1-\rho)} e^{-\frac{1}{1-\rho}\left(\frac{r_1^{\beta_2}}{\Omega_2}+\frac{r_2^{\beta_2}}{\Omega_2}\right)} I_0\left[\frac{2\sqrt{\rho}r_1^{\beta_2/2}r_2^{\beta_2/2}}{(1-\rho)\Omega_2}\right] + P_1 A(r_1,\beta_2,\Omega_2)A(r_2,\beta_1,\Omega_1) + P_2\left(1-e^{-\frac{r_t^{\beta_2}}{\Omega_2}}\right)\frac{\beta_1^{\,2}(r_1r_2)^{\beta_1-1}}{\Omega_1^{\,2}(1-\rho)} e^{-\frac{1}{1-\rho}\left(\frac{r_1^{\beta_1}}{\Omega_1}+\frac{r_2^{\beta_1}}{\Omega_1}\right)} I_0\left[\frac{2\sqrt{\rho}r_1^{\beta_1/2}r_2^{\beta_1/2}}{(1-\rho)\Omega_1}\right] + P_2 A(r_1,\beta_1,\Omega_1)A(r_2,\beta_2,\Omega_2) \quad (3)$$

For $r_1\geq r_t$, $r_2\geq r_t$

$$p_{r_1r_2}(r_1,r_2) = P_1\frac{\beta_1^{\,2}(r_1r_2)^{\beta_1-1}}{\Omega_1^{\,2}(1-\rho)} e^{-\frac{1}{1-\rho}\left(\frac{r_1^{\beta_1}}{\Omega_1}+\frac{r_2^{\beta_1}}{\Omega_1}\right)} I_0\left[\frac{2\sqrt{\rho}r_1^{\beta_1/2}r_2^{\beta_1/2}}{(1-\rho)\Omega_1}\right] + P_1 A(r_1,\beta_1,\Omega_1)\frac{\beta_2}{\Omega_2} r_2^{\beta_2-1} e^{-\frac{r_2^{\beta_2}}{\Omega_2}} + P_1 A(r_1,\beta_2,\Omega_2)A(r_2,\beta_1,\Omega_1) + P_1\left(1-e^{-\frac{r_t^{\beta_1}}{\Omega_1}}\right)\frac{\beta_2^{\,2}(r_1r_2)^{\beta_1-1}}{\Omega_2^{\,2}(1-\rho)} e^{-\frac{1}{1-\rho}\left(\frac{r_1^{\beta_2}}{\Omega_2}+\frac{r_2^{\beta_2}}{\Omega_2}\right)} I_0\left[\frac{2\sqrt{\rho}r_1^{\beta_2/2}r_2^{\beta_2/2}}{(1-\rho)\Omega_2}\right] + P_2\frac{\beta_1^{\,2}(r_1r_2)^{\beta_1-1}}{\Omega_1^{\,2}(1-\rho)} e^{-\frac{1}{1-\rho}\left(\frac{r_1^{\beta_1}}{\Omega_1}+\frac{r_2^{\beta_1}}{\Omega_1}\right)} I_0\left[\frac{2\sqrt{\rho}r_1^{\beta_1/2}r_2^{\beta_1/2}}{(1-\rho)\Omega_1}\right] + P_2 A(r_1,\beta_2,\Omega_2)\frac{\beta_1}{\Omega_1} r_2^{\beta_1-1} e^{-\frac{r_2^{\beta_1}}{\Omega_1}} + P_2 A(r_1,\beta_1,\Omega_1)A(r_2,\beta_2,\Omega_2) + P_2\left(1-e^{-\frac{r_t^{\beta_2}}{\Omega_2}}\right)\frac{\beta_1^{\,2}(r_1r_2)^{\beta_1-1}}{\Omega_1^{\,2}(1-\rho)} e^{-\frac{1}{1-\rho}\left(\frac{r_1^{\beta_1}}{\Omega_1}+\frac{r_2^{\beta_1}}{\Omega_1}\right)} I_0\left[\frac{2\sqrt{\rho}r_1^{\beta_1/2}r_2^{\beta_1/2}}{(1-\rho)\Omega_1}\right] + \quad (4)$$

where

$$A(r,\beta,\Omega) = \frac{\beta}{\Omega} r^{\beta-1} e^{-\frac{r^{\beta}}{\Omega}}\left[1-Q_1\left(\frac{\sqrt{2\rho}}{\sqrt{\Omega(1-\rho)}} r^{\beta/2}, \frac{\sqrt{2}}{\sqrt{\Omega(1-\rho)}} r_t^{\beta/2}\right)\right].$$

The outputs of SSC combiner are used as inputs for MRC combiner. The PDF at the output of SSC/MRC combiner with two branches is given by [18]:

$$p_r(r) = \int_0^r p_{r_1r_2}(r_1, r-r_1)dr_1 \quad (5)$$

Partials PDFs $p_i(r)$ for the SSC/MRC combiner output signal can be obtained substituting (1 - 4) in (5). For $r_1 < r_T$, $r - r_1 < r_T$ it is:

$$p_1(r) = \int_0^r dr_1 \left[P_1 A(r_1, \beta_2, \Omega_2) A(r_2, \beta_1, \Omega_1) + \right.$$
$$\left. + P_2 A(r_1, \beta_1, \Omega_1) A(r_2, \beta_2, \Omega_2) \right] \quad (6)$$

For $r_1 \geq r_T$, $r - r_1 < r_T$

$$p_2(r) = \int_{r_t}^r dr_1 \left[P_1 A(r_1, \beta_1, \Omega_1) \frac{\beta_2}{\Omega_2} r_2^{\beta_2 - 1} e^{-\frac{r_2^{\beta_2}}{\Omega_2}} + \right.$$
$$+ P_1 A(r_1, \beta_2, \Omega_2) A(r_2, \beta_1, \Omega_1) +$$
$$\left. + P_2 A(r_1, \beta_2, \Omega_2) \frac{\beta_1}{\Omega_1} r_2^{\beta_1 - 1} e^{-\frac{r_2^{\beta_1}}{\Omega_1}} + P_2 A(r_1, \beta_1, \Omega_1) A(r_2, \beta_2, \Omega_2) \right] \quad (7)$$

For $r_1 < r_T$, $r - r_1 \geq r_T$:

$$p_3(r) = \int_0^{r_t} dr_1 \left[P_1 \left(1 - e^{-\frac{r_t^{\beta_1}}{\Omega_1}} \right) \frac{\beta_2^2 (r_1 r_2)^{\beta_1 - 1}}{\Omega_2^2 (1-\rho)} e^{-\frac{1}{1-\rho}\left(\frac{r_1^{\beta_1}}{\Omega_2} + \frac{r_2^{\beta_2}}{\Omega_2} \right)} I_0 \left[\frac{2\sqrt{\rho} r_1^{\beta_2/2} r_2^{\beta_2/2}}{(1-\rho)\Omega_2} \right] + \right.$$
$$+ P_1 A(r_1, \beta_2, \Omega_2) A(r_2, \beta_1, \Omega_1) +$$
$$+ P_2 \left(1 - e^{-\frac{r_t^{\beta_2}}{\Omega_2}} \right) \frac{\beta_1^2 (r_1 r_2)^{\beta_1 - 1}}{\Omega_1^2 (1-\rho)} e^{-\frac{1}{1-\rho}\left(\frac{r_1^{\beta_1}}{\Omega_1} + \frac{r_2^{\beta_1}}{\Omega_1} \right)} I_0 \left[\frac{2\sqrt{\rho} r_1^{\beta_1/2} r_2^{\beta_1/2}}{(1-\rho)\Omega_1} \right] +$$
$$\left. + P_2 A(r_1, \beta_1, \Omega_1) A(r_2, \beta_2, \Omega_2) \right] \quad (8)$$

For $r_1 \geq r_T$, $r - r_1 \geq r_T$

$$p_4(r) = \int_{r_t}^r dr_1 \left[P_1 \frac{\beta_1^2 (r_1 r_2)^{\beta_1 - 1}}{\Omega_1^2 (1-\rho)} e^{-\frac{1}{1-\rho}\left(\frac{r_1^{\beta_1}}{\Omega_1} + \frac{r_2^{\beta_1}}{\Omega_1} \right)} I_0 \left[\frac{2\sqrt{\rho} r_1^{\beta_1/2} r_2^{\beta_1/2}}{(1-\rho)\Omega_1} \right] + \right.$$
$$+ P_1 A(r_1, \beta_1, \Omega_1) \frac{\beta_2}{\Omega_2} r_2^{\beta_2 - 1} e^{-\frac{r_2^{\beta_2}}{\Omega_2}} + P_1 A(r_1, \beta_2, \Omega_2) A(r_2, \beta_1, \Omega_1) +$$
$$+ P_1 \left(1 - e^{-\frac{r_t^{\beta_1}}{\Omega_1}} \right) \frac{\beta_2^2 (r_1 r_2)^{\beta_1 - 1}}{\Omega_2^2 (1-\rho)} e^{-\frac{1}{1-\rho}\left(\frac{r_1^{\beta_2}}{\Omega_2} + \frac{r_2^{\beta_2}}{\Omega_2} \right)} I_0 \left[\frac{2\sqrt{\rho} r_1^{\beta_2/2} r_2^{\beta_2/2}}{(1-\rho)\Omega_2} \right] +$$
$$+ P_2 \frac{\beta_1^2 (r_1 r_2)^{\beta_1 - 1}}{\Omega_1^2 (1-\rho)} e^{-\frac{1}{1-\rho}\left(\frac{r_1^{\beta_1}}{\Omega_1} + \frac{r_2^{\beta_1}}{\Omega_1} \right)} I_0 \left[\frac{2\sqrt{\rho} r_1^{\beta_1/2} r_2^{\beta_1/2}}{(1-\rho)\Omega_1} \right] +$$
$$+ P_2 A(r_1, \beta_2, \Omega_2) \frac{\beta_1}{\Omega_1} r_2^{\beta_1 - 1} e^{-\frac{r_2^{\beta_1}}{\Omega_1}} + P_2 A(r_1, \beta_1, \Omega_1) A(r_2, \beta_2, \Omega_2) +$$
$$\left. + P_2 \left(1 - e^{-\frac{r_t^{\beta_2}}{\Omega_2}} \right) \frac{\beta_1^2 (r_1 r_2)^{\beta_1 - 1}}{\Omega_1^2 (1-\rho)} e^{-\frac{1}{1-\rho}\left(\frac{r_1^{\beta_1}}{\Omega_1} + \frac{r_2^{\beta_1}}{\Omega_1} \right)} I_0 \left[\frac{2\sqrt{\rho} r_1^{\beta_1/2} r_2^{\beta_1/2}}{(1-\rho)\Omega_1} \right] + \right] \quad (9)$$

The PDF at the output of MRC/SSC combiner is the sum of components $p_i(r)$:

$$p_r(r) = p_1(r) + p_2(r) + p_3(r) + p_4(r) \quad (10)$$

Relatively simple closed form expressions for representing $p_r(r)$ can not be derived, because (6) – (9) are too complex for tractable communication system analyses, but the PDF can be evaluated numerically using software tools.

IV. OUTAGE PROBABILITY AND AMOUNT OF FADING

The outage probability is a very useful performance measure for diversity systems operating in fading environments. The outage probability is defined as the probability that the combiner output signal value falls below a given threshold r_{th}, also known as a protection ratio. The outage probability $P_{out}(r_{th})$ is defined as [17]:

$$P_{out}(r_{th}) = \int_0^{r_{th}} p_r(r) dr \quad (11)$$

Substituting (10) in (11), $P_{out}(r_{th})$ can be written as:

$$P_{out}(r_{th}) = \int_0^{r_{th}} \left[p_1(r) + p_2(r) + p_3(r) + p_4(r) \right] dr \quad (12)$$

Amount of fading (AF) is a unified measure of the severity of fading for particular channel model and is typically independent of the average fading power, but is dependent of the instantaneous SNR. Amount of fading for MRC combiner is defined by [17]:

$$AF = \frac{E[r^2]}{(E[r])^2} - 1 = \frac{E[(r_1 + r_2)^2]}{(E[(r_1 + r_2)])^2} - 1 \quad (13)$$

where $E(r)$ is N-th moment of r. By putting (6) – (9) into (13), AF is finally:

$$AF = \frac{\int_0^{\infty} \left[p_1(r) + p_2(r) + p_3(r) + p_4(r) \right] r^2 dr}{\left(\int_0^{\infty} \left[p_1(r) + p_2(r) + p_3(r) + p_4(r) \right] r dr \right)^2} - 1 \quad (14)$$

V. NUMERICAL RESULTS

The bit error rate curves, for different types of combiners and correlation parameters, are presented in Figs. 2 and 3.

It is assumed that both branches at the input have the same channel parameters. r_t is the optimal decision threshold. It is defined as [14]:

$$r_t = \Omega^{\frac{1}{\beta}} \Gamma\left(1+\frac{1}{\beta}\right) \quad (15)$$

The family of curves for the outage probabilities is given in Fig. 2 for four different cases: 1) one channel receiver, 2) MRC combiner at one time instant, 3) SSC/MRC combiner at two time instants for uncorrelated case and 4) SSC/MRC combiner at two time instants for very strong correlation.

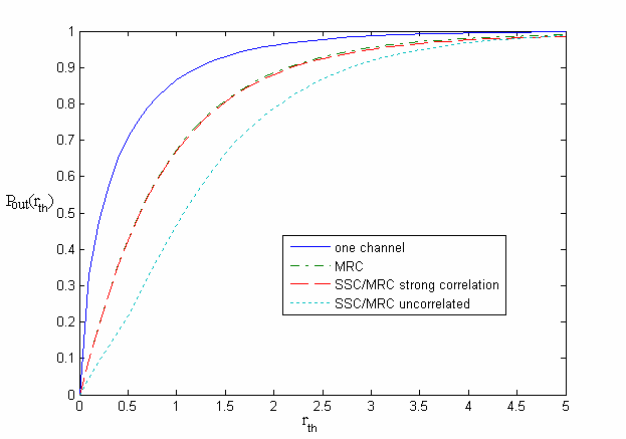


Figure 2. Outage probability for different types of combiners for parameters β=*0.7* and Ω=*0.5*

One can see from this figure that SSC/MRC combiner has significant better performance for both, uncorrelated case and for completely correlated signals (ρ =1), regarding classical MRC combiner at one time instant and one single channel receiver.

The outage probability curves for SSC/MRC combiner, for different values of correlation coefficient ρ, are presented in Fig. 3.

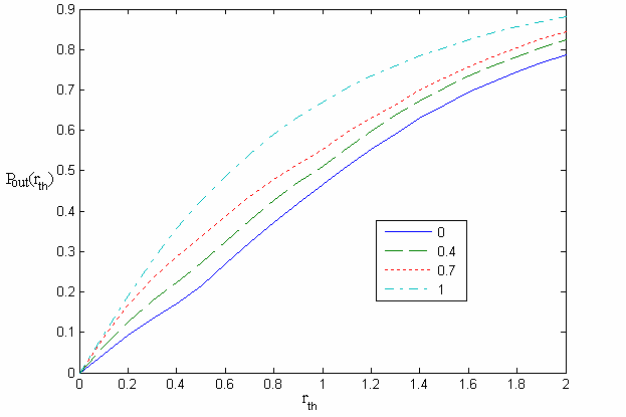


Figure 3. Outage probability for SSC/MRC combiner for parameters β=*0.7* and Ω=*0.5* and for different values of correlation coefficient ρ

It can be seen from Fig. 3 that SSC/MRC combiner has better performance for uncorrelated signals.

The amount of fading for one channel receiver, for MRC combiner at one time instant and for SSC/MRC combiner at two time instants, for uncorrelated case and for very strong correlation, is shown in Fig. 4.

It can be seen from Fig. 4 that SSC/MRC combiner has, once again, the best performance for uncorrelated case then the other combiners. The worst case is the case with one channel, i.e., without diversity combining at all.

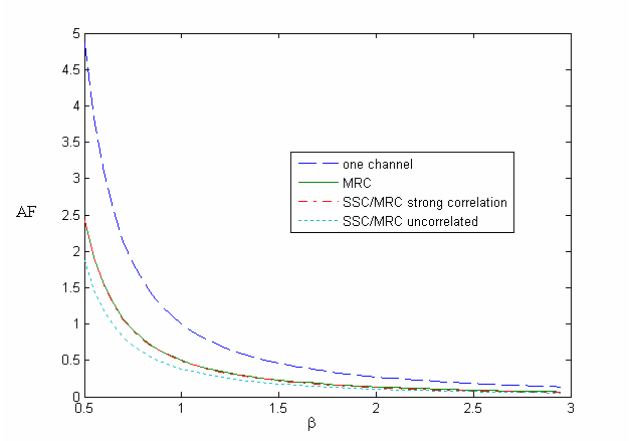


Figure 4. Amount of fading for different types of combiners for Ω=*0.5*

One can conclude from these figures that complex SSC/MRC combiner has better performance for uncorrelated case then MRC combiner at one time instant. For the value of correlation coefficient ρ = 1, the results of complex SSC/MRC combiner follows the results for MRC combiner. It is evident that usage of this complex SSC/MRC combiner gives better performance in the entire range, except in the case of strongly correlated signals. In this situaton it is not economic to use complex combiner. Thus, the advantages of utilization such type of complex combiner increases with decreasing of correlation between input signals.

VI. CONCLUSION

In diversity systems, SSC and MRC combining are often used techniques for combining signals. The performance of systems which make decision by two samples could be determined by the joint probability density function of SSC combiner output signals at two time instants. Then, they are involved in MRC combiner. Here, the joint PDF of SSC/MRC combiner output signal at two time instants in the presence of Weibull fading, derived earlier in closed form expressions, is used for outage probability and amount of fading determination.

The results are presented in some figures versus different parameters. It is evident that system performance is upgraded by use the mannerism of this complex combiner and sampling at two time moments. Obtained results are better compared with classical SSC and MRC combiners

 ISBN: 978-1-61208-279-0

except for very strong correlation between signals. In that case the performance curves for complex combiner and MRC combiner are coinciding.

ACKNOWLEDGMENT

This work has been funded by the Serbian Ministry for Science under the projects TR33035 and III 44006.

REFERENCES

[1] M. K. Simon and M. S. Alouini, Digital Communication over Fading Channels, Second Edition, Wiley-Interscience, A John Wiley&Sons, Inc., Publications, New Jersey, 2005.

[2] M. S. Alouini`and M. K. Simon, "Performance of generalized selection combining over Weibull fading channels", in Proc. IEEE Vehicular Technology Conference, Atlantic City, NJ, USA, May 2001, pp. 1735–1739.

[3] J. Cheng, C. Tellambura and N. C. Beaulieu, "Performance analysis of digital modulations on Weibull fading channels", in Proc. IEEE Vehicular Technology Conference, (VTC-Fall'03), Orlando, FL, Oct 2003, pp. 236–240.

[4] N. C. Sagias, G. K. Karagiannidis, P. S. Bithas and P. T. Mathiopoulos, "On the Correlated Weibull Fading Model and Its Applications", Proc. Vehicular Techology Conference IEEE, vol. 4, Sept. 2005, pp. 2149–2153.

[5] N. S. Adawi, et al., "Coverage prediction for mobile radio systems operating in the 800/900 MHz frequency range", IEEE Trans. Veh. Technol., vol. 37, Feb. 1988, pp. 3–72.

[6] H. Hashemi, "The indoor radio propagation channel", Proceedings IEEE 81, vol. 7, 1993, pp. 943–968.

[7] N. C. Sagias and G. K. Karagiannidis, "Gaussian class multivariate Weibull distributions: Theory and applications in fading channels", IEEE Transactions on Information Theory, vol. 51 (10), 2005, pp. 3608-3619.

[8] N. C. Sagias, D. A. Zogas, G. K. Karagiannidis and G. S. Tombras, "Channel capacity and second order statistics in Weibull fading", IEEE Communications Letters, vol. 8 (6), 2004, pp. 377-379.

[9] S. Khatalin and J. P. Fonseka, "On the channel capacity in Rician and Hoyt fading environments with MRC diversity," IEEE Trans. Veh. Technol., vol. 55, no. 1, Jan. 2006, pp. 137–141.

[10] Y. C. Ko, M. S. Alouini, and M. K. Simon, „Analysis and optimization of switched diversity systems"', IEEE Trans. Veh. Technol., vol. 49, Issue 5, Sept. 2000, pp.1813-1831.

[11] N. Sekulović, M. Stefanović, D. Drača, A. Panajotović and M. Zdravković, „Switch and stay combining diversity receiver in microcellular mobile radio system", Electrical Review (Przeglad Elektrotechniczny), vol. 86, no. 2, Dec. 2010, pp. 346-350.

[12] P. B. Nikolic, M. C. Stefanović, D. S. Krstić, and G. Lj. Stamenović, "The Joint Probability Density Function of the SSC Combiner Output in the Presence of Weibull Fading", XV International Symposium on Theoretical Electrical Engineering, ISTET'09, Lübeck, Germany, 22 – 24 June 2009, pp. 20-24.

[13] P. Nikolić, D. Krstić, M. Milić, and M. Stefanović, "Performance Analysis of SSC/SC Combiner at Two Time Instants in The Presence of Rayleigh Fading", Frequenz. Volume 65, Issue 11-12, Nov. 2011, pp. 319–325.

[14] M. Stefanović, P. Nikolić, D. Krstić, and V. Doljak, "Outage probability of the SSC/SC combiner at two time instants in the presence of lognormal fading", Przeglad Elektrotechniczny (Electrical Review), ISSN 0033-2097, R. 88 NR 3a/2012, March 2012, pp.237-240.

[15] D. Krstić, P. Nikolić and D. Radenković, "The Performances of Complex SSC/MRC Combiner in the Presence of Rayleigh Fading", Network Protocols and Algorithms, ISSN 1943-3581, 2012, Vol. 4, No. 3, DOI:10.5296/npa.v4i3.2055, URL: http://dx.doi.org/10.5296/npa.v4i3.2055, http://www.macrothink.org/journal/index.php/npa/article/view/2055/2112, , pp. 35-45

[16] D. Krstić, M. Stefanović and P. Nikolić, „Bit Error Rate for Complex SSC/MRC Combiner in the Presence of Nakagami-m Fading", Proc. of The Eighth Advanced International Conference on Telecommunications AICT 2012, ISBN: 978-1-61208-199-1, May 27 - June 1, 2012 - Stuttgart, Germany, pp. 75-80.

[17] D. Krstić, P. Nikolić, G. Stamenović and M. Stefanović, "The Bit Error Rate for Complex SSC/MRC Combiner at Two Time Instants in the Presence of Hoyt Fading", International Journal on Advances in Telecommunications, ISSN: 1942-2601, vol. 5, no. 1 & 2, year 2012, http://www.iariajournals.org/telecommunications/tele_v5_n12_2012_paged.pdf, pp. 69-78.

[18] D. Krstić, P. Nikolić, G.Stamenović and M. Stefanović, "Bit error rate for SSC/MRC Combiner at Two Time Instants in The Presence of log-normal Fading", Facta Universitatis. Series Automatic Control and Robotics, ISSN 1820-6417, Vol.10, No 1, 2011, UDC 621.396.94 621.395.38 519.724, pp. 83–95

[19] P. S. Bithas, G. K. Karagiannidis and N. C. Sagias, "Performance Analysis of a Class of GSC Receivers Over Nonidentical Weibull Fading Channels", IEEE Trans. Veh. Technol., vol. 54, Nov. 2005, pp. 1963–1970.

 ISBN: 978-1-61208-279-0

Iterative Detection of *M*-FSK Signal on MIMO Frequency Selective Fading Channels

Yuichi Yamane
Dept. of Computer Science and Engineering
Nagoya Institute of Technology
Nagoya, Japan
E-mail: 23417627@stn.nitech.ac.jp

Yasunori Iwanami
Dept. of Computer Science and Engineering
Nagoya Institute of Technology
Nagoya, Japan
E-mail: iwanami@nitech.ac.jp

Abstract— **In this paper, we have proposed the novel demodulation scheme of *M*-FSK (*M*-ary Frequency Shift Keying) signal on MIMO (Multiple Input Multiple Output) frequency selective channels. As for this demodulation scheme, there exists almost no investigation except a few reports. The proposed scheme uses FDE (Frequency Domain Equalization) and ISI (Inter-Symbol Interference) canceller plus MLD (Maximum Likelihood Detection). We further reduced the BER by using iterative feedback of detected results. The novelty in this paper is the integration of those techniques for detecting MIMO *M*-FSK signal. Through computer simulation, we have verified that the proposed scheme using FDE and ISI canceller plus MLD with iterative feedback exhibits the excellent BER characteristics compared with previously reported FDE detector.**

Keywords-MIMO, M-FSK, ISI, IAI, MLD, FDE, Multipath channel

I. INTRODUCTION

M-FSK signal has the constant envelope property and is appropriate to be amplified by nonlinear amplifier with high power efficiency. However, as *M*-FSK is a nonlinear modulation scheme, the equalization at the receiver side has been difficult when it is subjected to frequency selective channels. On the other side, due to the increasing demand of high data rate and reliable data transmission, MIMO schemes with multiple transmit and receive antennas become quite popular recently. The conventional MIMO scheme processes the received signals using linear matrix processing. However it has been difficult to apply the linear processing to the nonlinear modulation such as MIMO *M*-FSK, and accordingly there was almost no research on MIMO *M*-FSK signaling. So, we aimed to develop the MIMO *M*-FSK transmission scheme and examine its BER.

TABLE I LIST of PARAMETERS

M : Number of modulation level of FSK
n_T : Number of transmit antenna
n_R : Number of receive antenna
T_s : Symbol duration
KT_s : Maximum symbol delay time of multipath waves
$T_n = nT_s$: Block length for a FFT block
Δt : Sampling interval; $\Delta t = T_s / c$
c : Number of samples per a symbol
$N = cn$: Number of FFT samples
k : Time index of desired symbol time to be detected
$T_{CP} (= c_{CP} \Delta t)$: Cyclic Plefix length
I : Number of iterative feedbacks of block decision result
L : Number of delay paths

We had already shown that the FDE (Frequency Domain Equalization) scheme using CP (Cyclic Prefix) [1] is applicable to the signal separation and equalization of MIMO *M*-FSK signals [2]-[4], where the FDE is done before the demodulation process of *M*-FSK signal. This method was originally developed for SISO (Single Input Single Output) FSK signals [5]. In [6], we developed the detection scheme using ISI canceller and MLD to further improve the BER. In that scheme, the ISI's caused by the past symbols already detected were cancelled by ISI canceller, but the ISI cancellation caused by the future symbols and the separation of spatially multiplexed signals of MIMO transmission were achieved by MLD. This leads to the complexity of MLD of M^{Kn_T} where M is the modulation level (number of symbols) of *M*-FSK, KT_s the maximum symbol delay time of multipath waves, T_s the symbol duration and n_T the number of transmit antennas. As the complexity of MLD grows exponentially with the increase of K and n_T, the realization of this detector looks quite difficult even though using M-algorithm instead of MLD [6]. In this paper, in order to improve the BER characteristics of the FDE detector [2]-[4] and reduce the complexity of the detector in [6], we propose the novel demodulator structure in which the FDE is firstly done to obtain the tentative decision results. Using the tentative decision results, the ISI replicas due to the transmit symbols other than the desired symbol are cancelled form the receive signal. If the cancellation is perfect, we can obtain the receive signal as if only the desired symbol is transmitted. Then the MLD is applied to the receive signal to separate the IAI (Inter-Antenna Interference) of MIMO transmission. At this stage the output of MLD is regarded as the decision results of 0-th iteration. Then the 0-th decision results are fed back to the ISI canceller and again used as the tentative decision results for ISI cancellation. The ISI cancelled receive signal is then fed to the MLD again. This iterative processing is repeated and results in the better BER convergence. The basic structure of the above iteration receiver was developed in [7],[8] for the linear SC (Single Carrier)-FDE signals. In the proposed MIMO *M*-FSK demodulator, the complexity of MLD is proportional to M^{n_T} which does not depend on the ISI symbol length of K and this greatly reduces the complexity of detector enabling the detector for actual implementation.

 ISBN: 978-1-61208-279-0

This paper is organized as follows. In Section II, the FDE receiver for MIMO *M*-FSK is described and simulated. In Section III, the proposed detector with FDE and ISI canceller plus MLD is introduced. In Section IV, the BER results of proposed demodulator through computer simulation are shown. In Section V, the complexity of detector is described. The paper concludes with Section VI.

II. MIMO *M*-FSK DETECTION USING FDE

In Fig.1 we show the transmitter and receiver block diagram for MIMO *M*-FSK when using the FDE at the receiver. At the transmitter, the data bits are *M*-FSK modulated and CP is added to the transmit symbol block like Fig.2 with the block length of $T_n = nT_s$ where n is the number of symbols in a block and T_s is the symbol duration. From each antenna, the symbol block is transmitted. At the receiver, after the removal of CP, the received baseband I and Q signals are sampled at the sampling frequency of $f_s = 1/\Delta t$ where Δt is the sampling interval, $T_s = c\Delta t$ and c is the integer number. The value of c is taken large enough to satisfy the sampling theorem and not to make aliasing for the received analog I and Q signals.

The power spectral densities of continuous phase *M*-FSK signals are shown in Fig.3.

As a block length $T_n = nT_s$ contains cn samples, the number of FFT points becomes $N = cn$. The sampled complex discrete time signals $(I_p + jQ_p), p = 1, \cdots, N$ are then FFT-transformed and the FDE based on MMSE criterion is performed at each frequency point. As the result of FDE, the ISI compensation and the signal separation of spatially multiplexed signals are made simultaneously. Then the frequency domain samples are IFFT-transformed and the time domain discrete samples are obtained.

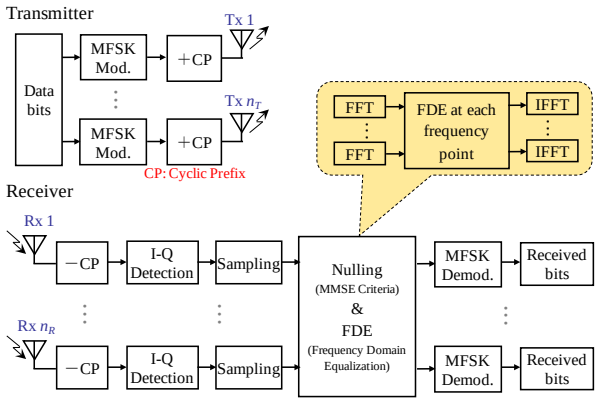


Figure 1. Block diagram of transmit and receive system of MIMO *M*-FSK using FDE at receiver on frequency selective MIMO channel

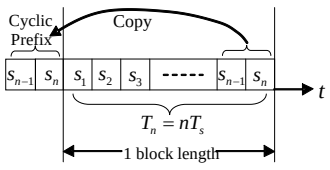


Figure 2. Insertion of CP at the transmitter

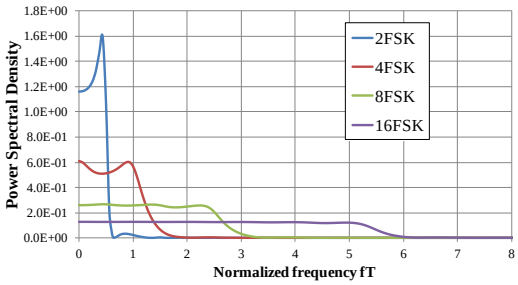


Figure 3. One-sided power spectral densities of *M*-FSK signals for modulation index h=0.7

These samples are equivalent to analogue *M*-FSK signals, because the sampling interval of Δt is taken small enough. The time domain samples are fed to the ordinary *M*-FSK detector. In this study, we employed a non-coherent energy detector as the *M*-FSK detector, because it is easy to implement and it does not need the phase synchronization [9]. In the energy detection of *M*-FSK signals, total *M* energy detectors are used, i.e., each energy detector for each frequency. The energy detectors for *M*-FSK are shown in Fig.4.

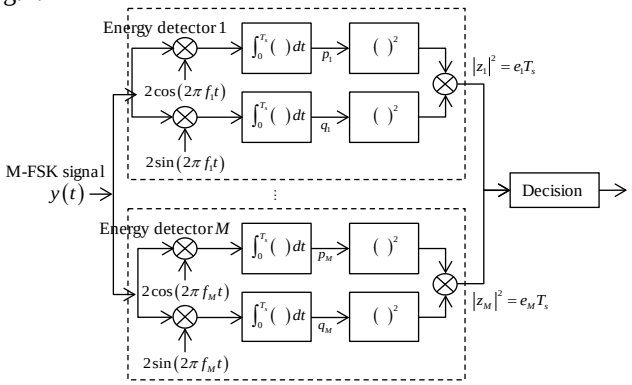


Figure 4. Energy detectors for *M*-FSK signal

The FDE weight with MMSE criterion is expressed as

$$\boldsymbol{D}(f) = E\{\boldsymbol{a}\boldsymbol{a}^H\}\boldsymbol{H}^H\left[\boldsymbol{H}E\{\boldsymbol{a}\boldsymbol{a}^H\}\boldsymbol{H}^H + \sigma^2\boldsymbol{I}_{n_R}\right]^{-1} \quad (1)$$

where $\boldsymbol{a} = \boldsymbol{a}(f)$ is the transmit signal vector at frequency point f, $\boldsymbol{H} = \boldsymbol{H}(f)$ the channel matrix, σ^2 the receive noise power, $E\{\ \}$ the ensemble average, $(\ )^H$ the Hermitian transpose, n_R the number of receive antenna, $\boldsymbol{I}_{n_R}$ the identity matrix. When $\boldsymbol{S}(f)$ denotes the power spectral density matrix of *M*-FSK signals, we have

$$E\{\boldsymbol{a}(f)\boldsymbol{a}^H(f)\} = \{\boldsymbol{S}(f)\Delta f\}\boldsymbol{I}_{n_T} = \{\boldsymbol{S}(f)[c/(NT_s)]\}\boldsymbol{I}_{n_T} \quad (2)$$

where n_T is the number of transmit antennas. By substituting (2) into (1), we obtain

$$\boldsymbol{D}(f) = \boldsymbol{H}^H\left[\boldsymbol{H}\boldsymbol{H}^H + \sigma^2 NT_s/\{c\boldsymbol{S}(f)\}\boldsymbol{I}_{n_R}\right]^{-1} \quad (3)$$

Using the FDE weight of (3), MIMO *M*-FSK signal is equalized and demodulated. The BER characteristics of FDE receiver for MIMO or SISO *M*-FSK with *M*=2, 4, 8 and 16 are examined through computer simulation. The computer simulation condition is shown in Table I. The delay profile between each transmit and receive antenna is illustrated in Fig.5. The BER results are shown in Fig.6.

From Fig.6, we observe that the BER characteristics are

 ISBN: 978-1-61208-279-0

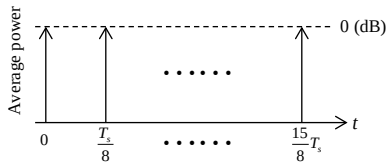


Figure 5. Power delay profile between each transmit and receive antenna

TABLE II SIMULATION CONDITION FOR MIMO *M*-FSK SIGNAL WITH FDE

Modulation	2 , 4 , 8 16 FSK
Modulation index	h=0.7
Number of Tx & Rx antennas	1×1 , 2×2 , 4×4
Channel model between each Tx and Rx antenna	Quasi-static 16 delay paths Rayleigh fading with equal power
Signal equalization and separation	FDE (MMSE criterion)
Symbol duration	T_s
Number of sample points in a symbol duration c	c=8,16,32,64 (2,4,8,16FSK) $T_s = c\Delta t$
Interval of delay paths	$T_s/8$
Maximum delay time	$15T_s/8$
Length of Cyclic Prefix	$2T_s$
Block length nT_s	$16T_s$
Number of FFT points (nc)	128, 256, 512, 1024

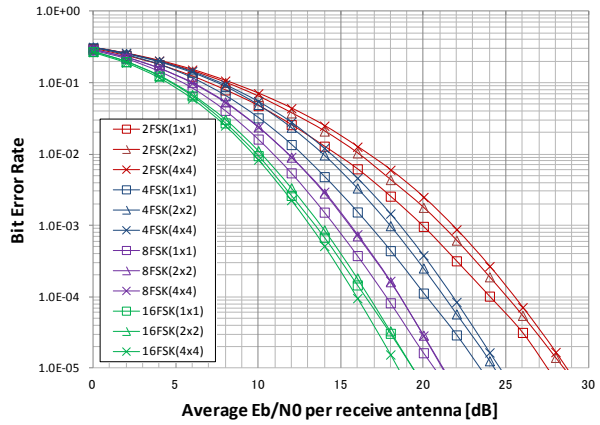


Figure 6. BER characteristics of MIMO *M*-FSK with FDE and energy detector

improved as the value of M increases from 2FSK, 4FSK, 8FSK to 16FSK. This improvement is due to the bandwidth expansion of *M*-FSK signal with larger M value. Due to the FDE, the diversity order, i.e., the gradient of BER curve, is the same among 1×1 , 2×2 and 4×4 . This is also observed in MIMO OFDM with FDE. Accordingly, we can say that the multi-stream MIMO transmission is available for *M*-FSK through the FDE receiver.

III. ITERATIVE DETECTION OF MIMO *M*-FSK USING FDE AND ISI CANCELLER PLUS MLD

By using the equalization and signal separation with FDE and energy detector, the demodulation of MIMO *M*-FSK signal can be achieved. However, the BER characteristic of FDE receiver is not enough and we aim to obtain further BER improvement. In order to do this, we consider the decision results from FDE as the tentative decision results. Using the tentative decision results, the ISI replica at the receiver is generated and is subtracted from the receive signal. If the tentative decision results from FDE are correct, we can cancel the ISI and obtain the receive signal as if only transmit symbols at desired time k are transmitted from n_T transmit antennas. For this receive signal, the separation of spatially multiplexing is done using MLD. Then we consider the outputs of MLD as the reliability enhanced decision results and use them as the evolved tentative decision results. Using the evolved results, the new ISI replicas are generated and again subtracted from the receive signal to cancel the ISI. After this ISI cancellation, the MLD is again employed to separate the spatial multiplex. This iterative processing is repeated to improve the BER for each iteration. The above iterative procedure algorithm was originally reported in [7] for MIMO SC-FDE receiver.

In Fig.7, the transmitter and receiver block diagram is shown for this iterative demodulator. At the transmitter, *M*-FSK symbols with the block length of nT_s are generated and the CP with the length of $c_{CP}\Delta t$ is added at the head of the block. Each block with CP is transmitted from each antenna. At the receiver, after removing the CP, the equalization and signal separation are firstly done through FDE and the tentative decision results are obtained. Those tentative decision results are fed to the ISI canceller and the replica for ISI cancellation is generated. The replicas are made in order to cancel the ISI's caused by the total $2K$ transmit symbols before and after the desired symbol at time k . Accordingly, in making the ISI replica, the transmit symbol at time k in the tentative decision results is set to zero. By subtracting the ISI replica from the receive signal, the ISI components due to the transmit symbols before and after the transmit symbol at time k are cancelled. If the tentative decision results are correct, we can get the receive signal as if the symbols at time k are only transmitted from antenna $1\sim n_T$ with spatially multiplex. Then in order to separate the spatial multiplex, the MLD is applied to the ISI cancelled receive signal and the decision results at time k are obtained. The new decision results at time k are sequentially fed back to the ISI canceller and the decision time is evolved from k to $(k+1)$. After obtaining all the decision results for n symbols, those decision results for n symbols are replaced

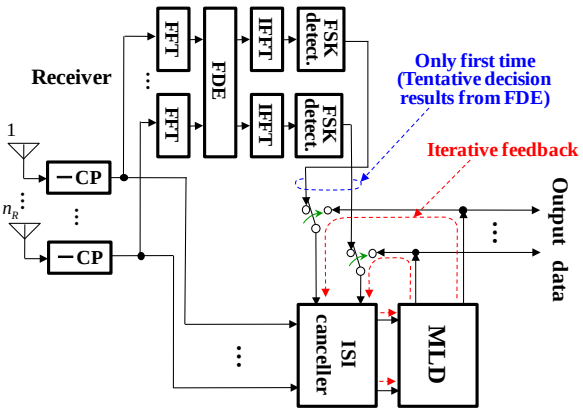


Figure 7. Iterative demodulation scheme of MIMO *M*-FSK using FDE and ISI canceller plus MLD (Transmitter is the same as Fig.1.)

 ISBN: 978-1-61208-279-0

as the new tentative decision results of a block. The above procedure for a FDE block with n symbols is repeated I times to obtain the final decision results.

We consider the ISI replica generation at the receiver for the first symbol ($k=1$) in a block with the length of nT_s. When the MLD is done 0 times, the ISI cancellation replica $\boldsymbol{y}'^{(0)}_{1,ij}$ for the 1st symbol in a block from transmit antenna j to receive antenna i is given by

$$\boldsymbol{y}'^{(0)}_{1,ij} = \begin{pmatrix} y^{(0)}_{1,ij}(c-1+c_{CP}) \\ \vdots \\ y^{(0)}_{1,ij}(0) \end{pmatrix}$$

$$= \begin{pmatrix} h_{ij}(0) & \cdots & h_{ij}(c_{CP}) & 0 & \cdots & 0 \\ 0 & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & h_{ij}(0) & \cdots & h_{ij}(c_{CP}) \end{pmatrix} \begin{pmatrix} x^{(0)}_j(c-1+c_{CP}) \\ \vdots \\ x^{(0)}_j(c) \\ 0 \\ \vdots \\ 0 \\ x^{(0)}_j(-1) \\ \vdots \\ x^{(0)}_j(-c_{CP}) \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \} c_{CP} \\ \\ \\ \} c \\ \\ \\ \} c_{CP} \\ \\ \end{matrix} \quad (4)$$

where the number of samples in a symbol T_s is c ($T_s = c\Delta t$), the CP length $c_{CP}\Delta t$, the sample of transmit signal $x^{(0)}_j(p)$ at $t = p\Delta t$ (p: integer number), the complex gain of l-th ($l = 0,1,\cdots,c_{CP}-1$) delay path $h_{ij}(l)$. Also the subscript "1" in $\boldsymbol{y}'^{(0)}_{1,ij}$ means the 1st symbol and the superscript "(0)" the execution times of MLD. In (4), the sample values at sample time $t = 0,\cdots,(c-1)\Delta t$ of the 1st symbol in a transmit block is set to zero. Also in (4), $x^{(0)}_j(-c_{CP}),\cdots,x^{(0)}_j(-1)$ and $x^{(0)}_j(c),\cdots,x^{(0)}_j(c-1+c_{CP})$ show the c_{CP} samples before and after the symbol at time $k=1$ respectively. As shown in Fig.8, the first part with the length of $T_{CP}(=c_{CP}\Delta t)$ and the last part of T_{CP} in a FFT block are circularly copied after the tail and before the head of a block respectively. This is because the FDE for obtaining the tentative decision results is based on the circular convolution in a time domain, i.e., we have to consider the ISI components falling in the first part T_{CP} in a FFT block and the ones falling in the subsequent part T_{CP} after the tail of a block.

The receive replica $\boldsymbol{y}''^{(0)}_{1,ij}$ for MLD of the 1st symbol in a block is given by

$$\boldsymbol{y}''^{(0)}_{1,ij} = \begin{pmatrix} y''^{(0)}_{1,ij}(c-1+c_{CP}) \\ \vdots \\ y''^{(0)}_{1,ij}(0) \end{pmatrix}$$

$$= \begin{pmatrix} h_{ij}(0) & \cdots & h_{ij}(c_{CP}) & 0 & \cdots & 0 \\ 0 & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & h_{ij}(0) & \cdots & h_{ij}(c_{CP}) \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ x^{(0)}_j(c-1) \\ \vdots \\ x^{(0)}_j(0) \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \\ \\ \} c_{CP} \\ \\ \\ \\ \} c_{CP} \\ \\ \end{matrix} \quad (5)$$

$\boldsymbol{y}''^{(0)}_{1,ij}$ in (5) is the output from the channel when only the 1st transmit symbol T_s is transmitted. In generating MLD replica, all the transmit symbols other than at time k are set to zero. Those two replicas in (4) and (5) have to be generated so as to satisfy the phase continuity of FSK, because we are assuming the continuous phase M-FSK.

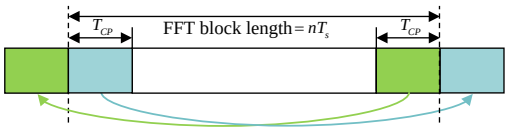


Figure 8. Termination processing in a FFT block for ISI cancellation

Using those receive replicas, the squared Euclidian distance between the ISI subtracted signal from the receive signal $\boldsymbol{y}_{1,i}$ at receive antenna i and the MLD replica is calculated. The candidate $\boldsymbol{x}_1^{(0)}$ of transmit symbol vector for the 1st symbol in a FFT block is determined so as to minimize the distance metric as shown below.

$$\boldsymbol{x}_1^{(0)} = \arg\min_{\boldsymbol{x}_1^{(0)}} \left\{ \sum_{i=1}^{n_R} \left\| \left(\boldsymbol{y}_{1,i} - \sum_{j=1}^{n_T} \boldsymbol{y}'^{(0)}_{1,ij} \right) - \sum_{j=1}^{n_T} \boldsymbol{y}''^{(0)}_{1,ij} \right\|^2 \right\} \quad (6)$$

where

$$\boldsymbol{x}_1^{(0)} = \left[\boldsymbol{x}^{(0)}_{1,1} \quad \cdots \quad \boldsymbol{x}^{(0)}_{1,n_T} \right]^T, \quad \boldsymbol{x}^{(0)}_{1,j} = \left[x^{(0)}_{1,j}(0) \quad \cdots \quad x^{(0)}_{1,j}(c-1) \right]^T \quad (7)$$

$$\boldsymbol{y}_{1,i} = \begin{pmatrix} y_i(c-1+c_{CP}) \\ \vdots \\ y_i(0) \end{pmatrix}$$

$$= \sum_{j=1}^{n_T} \begin{pmatrix} h_{ij}(0) & \cdots & h_{ij}(c_{CP}) & 0 & \cdots & 0 \\ 0 & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & h_{ij}(0) & \cdots & h_{ij}(c_{CP}) \end{pmatrix} \begin{pmatrix} x^{(0)}_j(c-1+c_{CP}) \\ \vdots \\ x^{(0)}_j(-c_{CP}) \end{pmatrix} \quad (8)$$

The decision result $\boldsymbol{x}_1^{(0)}$ of MLD in (6) is regarded as the new tentative decision result and is fed back to the ISI canceller to detect the 2nd symbol in a block at time $k=2$. Next, for the 2nd symbol, the ISI cancellation and MLD are executed and the decision result of $\boldsymbol{x}_2^{(0)}$ is obtained. Then $\boldsymbol{x}_2^{(0)}$ is regarded as the new tentative decision result and is fed back to the ISI canceller for determining the 3rd symbol $\boldsymbol{x}_3^{(0)}$. Those sequential feedback procedure is repeated up to the n-th symbol in a block. The obtained decision results for all the n symbols in a block are regarded as the block decision results, which is again fed back to the ISI canceller and the subsequent MLD processing is done. This feedback of block decision results is repeated I times and the final decision results are obtained.

IV. SIMULATION RESULTS OF ITERATIVE DETECTION RECEIVER FOR MIMO M-FSK

Computer simulations are made to verify the BER improvement of iterative receiver. The simulation condition is listed in Table III. The delay profile is the same as in Fig.5. The BER results are shown in Fig.9 to Fig.14.

From Fig.9~14, we observe that the BER characteristics of "FDE and ISI canceller plus MLD" receiver with iterative detection are better than "FDE" receiver. We also see the BER is improved as the number I of iterative feedbacks increases. $I=4$ (#4) is enough for the BER convergence, where the MLD is repeated five times for a block.

 ISBN: 978-1-61208-279-0

V. COMPARISON OF COMPLEXITY OF RECEIVER STRUCTURE

We compared the complexity between the receiver with "FDE" and the one with "FDE and ISI canceller plus MLD." The complexity is compared as the number of complex additions and multiplications required for detecting one symbol of FSK. The equations for calculating the complexity are given in Table IV, where we assume $I=1$, the maximum delay time of multipath wave is $(L-1)(2\Delta t)$ and $O\left(n_T{}^2\right)=n_T{}^2$. We also show the numerical results of complexity in Fig.15 and Fig.16. From those results, we know that the complexity of FDE receiver is far less than the "FDE and ISI canceller plus MLD" receiver. For the FDE receiver, the complexity gradually increases as the number of symbols n in a FFT block becomes large, but does not depend on the delay time parameter L of multipath. On the other hand, for the "FDE and ISI canceller plus MLD" receiver, we observe that the complexity increases as L becomes large, but does not depend on the number of symbols n in a block.

TABLE III SIMULAION CONDITIONS FOR MIMO *M*-FSK WITH ITERATIVE DETECTION

Modulation	2 , 4 , 8 FSK
Modulation index	h=0.7
Number of Tx & Rx antennas	2×2 , 4×4
Channel model between each Tx and Rx antenna	Quasi-static 16 delay paths Rayleigh fading with equal power
Signal equalization and separation	FDE (MMSE criterion) & ISI canceller + MLD
Symbol duration	T_s
Number of sample points for symbol duration c	c=8,16,32 (2,4,8FSK) $T_s=c\Delta t$
Interval of delay paths	$T_s/8$
Maximum delay time	$15T_s/8$
Length of Cyclic Prefix	$2T_s$
Block length nT_s	$16T_s$
Number of FFT points (nc)	128, 256, 512
Number of block iteration I	#0, #1, #2, #3, #4

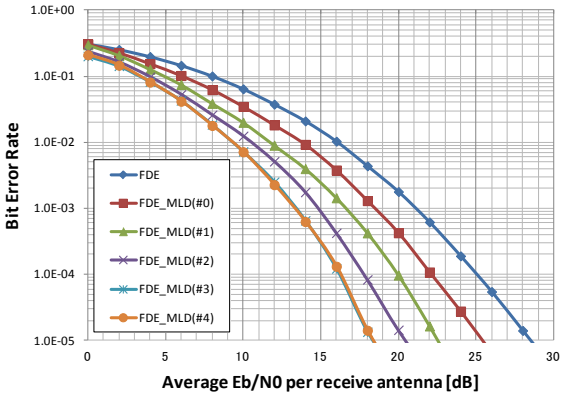


Figure 9. BER comparison between conventional FDE receiver and proposed iterative receiver (2FSK, 2×2)

VI. CONCLUSIONS

In this study, we have proposed the new receiver structure for *M*-FSK signal on frequency selective MIMO channels. In addition to the FDE receiver which was proposed previously, we demonstrated the novel receiver structure in which the ISI components are firstly cancelled by the tentative decision results obtained by FDE and then the MLD is employed to separate the spatial multiplexing.

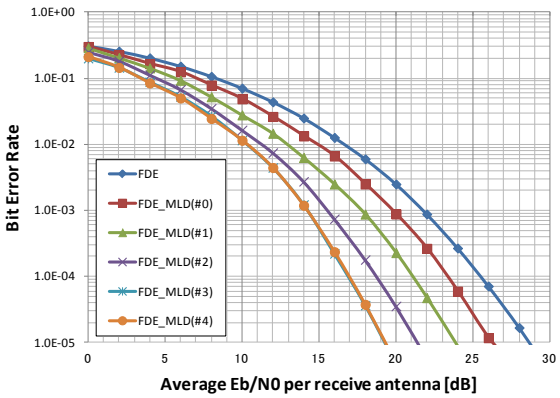


Figure 10. BER comparison between conventional FDE receiver and proposed iterative receiver (2FSK, 4×4)

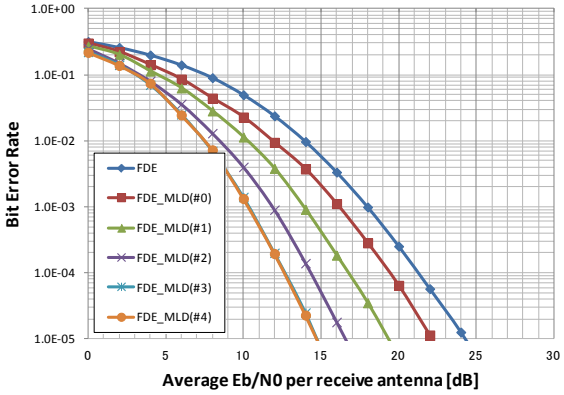


Figure 11. BER comparison between conventional FDE receiver and proposed iterative receiver (4FSK, 2×2)

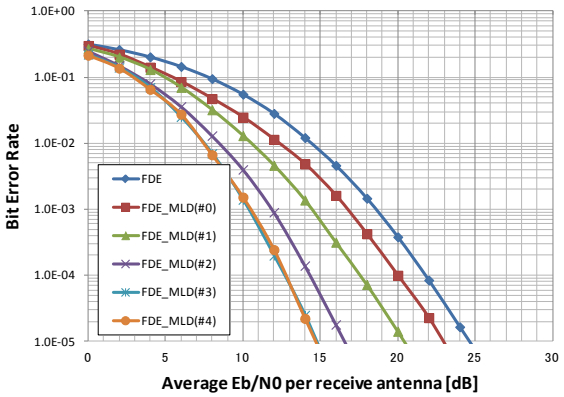


Figure 12. BER comparison between conventional FDE receiver and proposed iterative receiver (4FSK, 4×4)

 ISBN: 978-1-61208-279-0

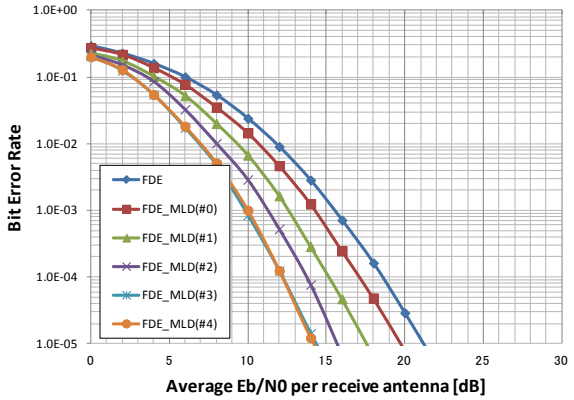


Figure 13. BER comparison between conventional FDE receiver and proposed iterative receiver (8FSK, 2×2)

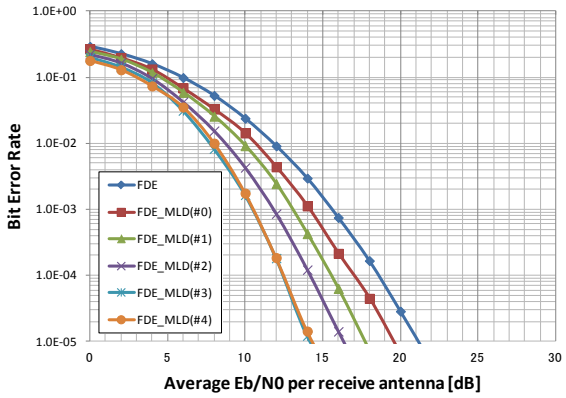


Figure 14. BER comparison between conventional FDE receiver and proposed iterative receiver (8FSK, 4×4)

TABLE IV NUMBER OF COMPLEX ADDITIONS AND MULTIPLICATIONS FOR DETECTING ONE *M*-FSK SYMBOL

FDE	$c\times\left[10\log_2(c\times n)+2\times O\left(n_T^{\,2}\right)+2n_T-13\right]$
FDE & ISI canceller + MLD	$c\times\left[10\log_2(c\times n)+2\times O\left(n_T^{\,2}\right)+2n_T-13\right]$ $+2\left\{c\times n_R\times(L\times n_T+1)\right\}\times\left(M^{n_T}+1\right)/n_T$

The proposed receiver improves the BER with the iterative feedback of decision results. The BER characteristics of proposed iterative receiver are improved very much when compared with the conventional FDE receiver. By using the tentative decision results obtained from the FDE, the complexity of MLD in the proposed iterative receiver does not depend on the transmit block length and becomes modest, while the high quality separation of spatially multiplexed signals that comes from MLD is maintained.

As future studies, instead of known channel state information (CSI) at the receiver, the measured CSI will be employed and the features of MIMO *M*-FSK comparing with existing linear modulations should be clarified.

ACKNOWLEDGEMENT

This study is partially supported by the Grants-in-Aid for Scientific Research 24560454 of the Japan Society for the Promotion of Science and the Sharp Corporation.

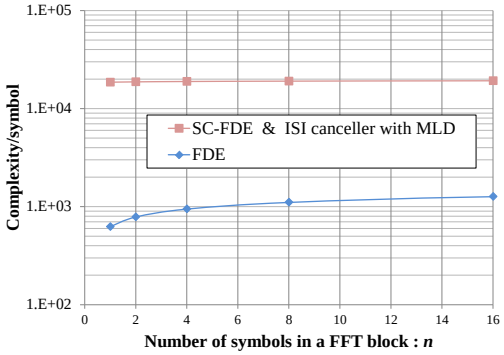


Figure 15. Complexity comparison of receivers for symbols *n* in a block

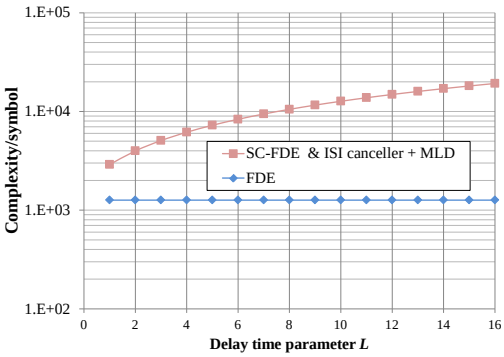


Figure 16. Complexity comparison of receivers for the maximum delay time of $(L-1)(2\Delta t)$

REFERENCES

[1] D. Falconer, S. L. Ariyavisitakul, A. Benyamin-Seeyar, and B. Eidson, "Frequency Domain Equalization for Single-Carrier Broadband Wireless Systems," IEEE Commun. Mag., April 2002, pp. 58-66.

[2] M. Hijimoto, Y. Iwanami, and E. Okamoto, "A study on M-ary FSK with non-coherent detection on MIMO frequency selective channels," Technical report of IEICE, RCS2008-230, Mar. 2009 (in Japanese), pp. 107-112.

[3] K. Nakayama, Y. Iwanami, and E. Okamoto, "A comparative study of MIMO-MFSK with frequency and energy detection under frequency-selective fading channels", IEICE Society Conference, Sep. 2009 (in Japanese), B-5-34.

[4] K. Nakayama, Y. Iwanami, and E. Okamoto, "MIMO MFSK receivers using FDE and MLD on quasi-static frequency selective fading channels," International Symposium on Information Theory and its Applications 2010 (ISITA2010), Taichung Taiwan, ES2-Mo-1, Oct. 2010, pp. 31-36.

[5] M. Maki and Y. Akaiwa, "An adaptive equalizer for FSK frequency detection," IEICE Technical Report, RCS 93-54, September 1993 (in Japanese), pp. 23-28.

[6] Y. Iwanami and K. Nakayama, "MLD-based MFSK Demodulation on MIMO Frequency Selective Fading Channel," The Seventh International Conference on Wireless and Mobile Communications, ICWMC2011 Luxemburg, June 2011, pp. 30-35, ISBN 978-1-61208-140-3.

[7] Y. Nouda, T. Koike; S. Yoshida, "Iterative MLD equalizer preceded by MIMO-FDE for wideband spatial multiplexing systems,"IEEE VTC2005 Spring, Vol.1, 30 May - 1 June 2005, pp. 533-537.

[8] M. Utsunomiya, Y. Iwanami and E. Okamoto, "An iterative signal detection scheme for MIMO SC-FDE using ISI canceller and MLD," IEICE Technical report, WBS2007-77, Feb. 2008 (in Japanese), pp. 89-94.

[9] Y. Iwanami and P. H. Wittke, "Error Performance Analysis of an Energy Sequence Estimation Receiver for Binary FSK on Frequency-Selective Fading Channels, IEEE Trans. on Wireless Communications, Vol. 2, No. 2, March 2003, pp. 260-269.

 ISBN: 978-1-61208-279-0

A wireless mesh network solution based on WiMAX technology with smart antennas support

Şerban Georgică Obreja,
University POLITEHNICA of Bucharest
Bucharest, Romania
e-mail: serban@radio.pub.ro

Alexey Baraev
Create-Net
Trento, Italy
e-mail: alexey.baraev@create-net.org

Irinel Olariu,
University POLITEHNICA of Bucharest
Bucharest, Romania
e-mail: irinel.olariu@elcom.pub.ro

Eugen Borcoci
University POLITEHNICA of Bucharest
Bucharest, Romania
e-mail: eugen.borcoci@elcom.pub.ro

***Abstract*—Wireless Mesh Networks represent a good solution to offer Internet access in sparse population areas. The latest wireless technologies, such as WiMAX, offer high data rates services, but they are not designed for wireless mesh architectures. This paper proposes a wireless mesh solution based on WiMAX nodes equipped with smart antennas, and presents some tests for performance evaluation of the proposed solution. The WiMAX mesh node is emulated by interconnecting, via Ethernet, a Base Station and a Subscriber Station. To avoid interference between the two radio interfaces, they are configured to operate on different frequency bands. The WiMAX mesh node is simulated in OPNET and its performance is compared with a real life implementation of such a hybrid mesh node. The performance evaluation was done using an experimental platform, consisting of both a simulated and a real life part, which are interconnected through the System in the Loop OPNET function.**

Keywords-mesh network; WiMAX; Smart Antennas; OPNET simulator; System in the Loop.

I. INTRODUCTION

Worldwide Interoperability for Microwave Access (WiMAX) was developed for broadband wireless networks, and it offers high data rate over long distances [1]. An application suitable for WiMAX technology is its usage in wireless backhaul networks, to provide coverage over low density areas or as alternative solutions to wired networks. Such a solution offers resilience in disaster scenarios obtained through self healing property facilitated by the wireless communication. [2].

Wireless Mesh Networks are the solution for wireless backhaul (WMN). On one side, the WMNs share common features with ad hoc networks (MANET), on the other side, the WMNs, being used for backbone, have minimal mobility and no constraint on power consumption. The WMN routing protocols can be generally adapted from those used in fixed networks and also from ad hoc networks, but, due to the specific conditions of WMN routers (limited mobility and no power constraints), the protocols like in fixed networks can be more appropriate.

The work presented in this paper was done in the framework of the SMART-Net FP7 project, which aimed to investigate the use of smart antennas in Wireless Mesh Networks, mainly based on WiMAX technology. A solution for WiMAX mesh networks based on hybrid WiMAX mesh nodes is investigated. This hybrid node was obtained by linking a Base Station with a Subscriber Station via Ethernet. Both stations include layer three capabilities, mainly routing capabilities. A node with relay properties is obtained, which has also routing capabilities. This solution was implemented and simulated in the OPNET simulator. A similar approach was developed by Thales Company for the real implementation of such a hybrid node [3]. Here, the Subscriber Station functionalities were implemented on a Base Station platform. Special timing synchronization between the uplink and downlink intervals for Base Station and Subscriber Station was introduced, in order to avoid interference between the two wireless interfaces, by not allowing them to transmit and receive simultaneously.

There are several works on wireless mesh networks based on WiMAX, but they focus on multihop relay networks built with WiMAX relay stations, built based on the IEEE 802.16j [1]. A spanning tree rooted at Base Station is built, with Subscriber Stations as leaf nodes. Usually, these works approach scheduling and channel assignment algorithms to optimize the resource allocation in multihop relay networks [4][5]. There are also approaches which consider multihop relay networks with smart antenna support, and which try to find solutions for joint resource allocation and beam scheduling in order to reduce the interference and increase the overall throughput [6][7]. Most of the proposed algorithms are based on centralized solutions. In this paper, due to the way the WiMAX mesh node is build, the solution is completely distributed, without any correlation between the scheduling or channel assignment at each node.

This paper is organized as follows. The second section is dedicated to Smart-Net system presentation. The third

 ISBN: 978-1-61208-279-0

section presents the experimental platform setup, while the fourth section presents the test scenarios and the results. The last section contains the conclusions and some suggestions for future work.

II. SMART-NET SYSTEM COMPONETS

A. SMART-Net system

As it was mentioned in the previous section, this solution was proposed in the SMART-Net project, which focused on introducing the smart antennas support in WiMAX technology and on developing a wireless mesh network for backhaul access, based on WiMAX nodes with multiple radio interfaces. To validate it, an experimental platform was developed during the project [8]. This test platform consists of a simulated part and a part realized with real WiMAX nodes, equipped with smart antennas. The WiMAX equipments used in the testbed are produced by Thales Company, while the smart antennas are produced by Plasma Antennas Company. Both are members of the Smart-Net project. The simulation platform was developed using the OPNET network simulator. The smart antennas were modeled in the OPNET, based on the real antennas measurement data, and integrated with the simulated WiMAX nodes. Also, the beam selection and tracking algorithms were implemented in OPNET and included in the WiMAX radio pipeline stages for the Smart Antennas control. The interconnection of the two test platform components was realized using the System-in-the-Loop (SITL) function provided by OPNET. The real part was developed at France Telecom premises located in Lannion, France, while the simulated part was developed at University Politehnica of Bucharest (UPB) in Romania. The two testbeds were interconnected through a VPN tunnel via GEANT network. The interconnection setup is illustrated in Fig. 1. With such an approach, the capability of the proposed backhaul network to operate over large areas was tried to be proved.

Some initial results of this work are presented in [9]. There, only the simulated network was evaluated. In this paper, both simulated and real network were integrated in an experimental platform. A similar approach is presented in [10], where SITL is used to evaluate WiFi wireless networks performances with real time traffic injected in the simulated network.

B. Smart Antenna features

One of the main topics of SMART-Net project was the introduction of smart antenna support in WiMAX. Smart antennas define antennas which are capable of adapting the radiation and reception pattern automatically. They provide a higher gain on the main beam direction and much smaller gains on the secondary beams, which will ensure a lower interference noise.

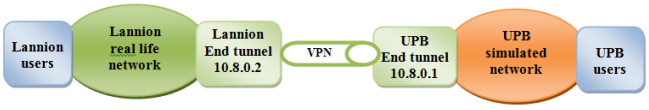


Figure 1. Testbeds interconnection setup

In practice, one can have two types of smart antennas. One is represented by adaptive smart antennas, which are capable of adapting its beam in every direction, by adaptively combining the signal from multiple antenna elements. They have the advantage of being able to adapt the beam in every direction, but with the drawback of higher switch time and increased complexity. The second type is represented by switched beam antennas, which has several fixed beams and can be selected individually. They have only a limited number of directions which can be obtained, but they are rapid and with a simpler implementation.

For the SMART-Net project, two types of switched multi-beam antennas, capable of WiMAX operation, have been designed and implemented [3] [11]. An active, 12 beam cylindrical array antenna with omnimode and a passive 9 beam planar array antenna with sectoral mode. The active 12-beam cylindrical antenna with 360° coverage is suitable for mesh and nomadic Point to Multipoint operation. It has typical ranges of up to 20 km, depending on the modulation rate. The passive 9 beam planar antenna, has narrower beams, and is suited for medium range backhaul and relay operations. A representation of both antennas in OPNET has also been provided as simulated data.

A switching algorithm is used to choose the appropriate beam. This algorithm is based on a learning interval in which, based on SINR, the best beam is chosen for each destination. Based on the decision taken by the selection algorithm, when a smart node (a node equipped with smart antennas) needs to communicate with another smart node, the beam with the best SINR is used. Because the best beam is decided in the learning phase, the switch operation is very fast, a few nanoseconds.

III. EXPERIMENTAL PLATFORM SETUP

The experimental platform infrastructure consists of a simulated WiMAX network, which is interconnected with real devices in order to introduce real time traffic in the simulation, Fig.1.

The System-in-the-Loop OPNET function allows real time traffic to be exchanged between real and simulated parts of the networks during the simulation. The actual OPNET version works only with an Ethernet network adaptor for SITL interconnection. The requirement of using Ethernet link between the real devices and the SITL gateway introduces limitation in developing joint real and simulated wireless network scenarios. For the SITL scenarios, the simulation must run in real-time. This requirement introduces an additional limitation for the SITL scenarios: it limits the number of nodes, which can be used in the simulation.

A. Real Life testbed infrastructure

The Real Life testbed was developed at the France Telecom Lannion premises. Its structure is presented in Fig. 2. Its main part consists of a WiMAX network with one Base Station, three Subscriber Stations and one Relay Station. The Relay Station [3] is built by combining a Subscriber Station and a Base Station together and synchronizing them in terms of uplink and downlink mapping: the scheduling for the

 ISBN: 978-1-61208-279-0

uplink and downlink transmission intervals for both WiMAX links, given by the UL and DL MAP fields from the WiMAX frame generated by the main Base Station and the Relay Base Station, must be synchronized in order to reduce the interference at the relay node.

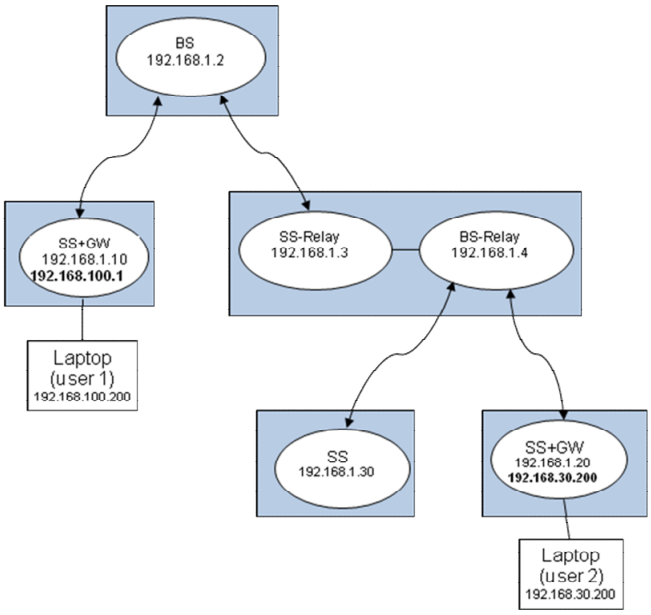


Figure 2. The real life testbed infrastructure

The WiMAX nodes were equipped with two antennas: the smart antenna having a beam gain of 16.5 dBi (beam mode) / 12.5 dBi (omni mode) and, for comparison reasons, a classic sector antenna of 17 dBi. They worked in the 5470 - 5725 MHz band at constant Equivalent Isotropic Radiated Power (EIRP) of 30 dBm. Consequently, the transmission power of the base station was set to 24 dBm and those of the subscriber node to 14 dBm (smart antenna beam mode) or 18 dBm (smart antenna sector mode). More details about the performance of the Lannion real life WiMAX network can be viewed in [12][13].

The WiMAX BS is connected with the Lannion Internet gateway. This gateway is used to host the VPN client which in used to connect to the VPN server located at UPB premises. The VPN tunnel is a layer 3 tunnel. From a logical point of view the Lannion test bed consists of a direct WiMAX link (a BS-SS link) and a relay based link (BS-SS_relay-BS_relay-SS) – Fig. 2. A stream flowing through the real life testbed will cross one or two WiMAX links, depending where it generates/ends (user 1 or user 2).

B. *Simulation testbed infrastructure*

For the simulated network the main WiMAX physical parameters are: 20MHz bandwidth, 2048 subcarriers, 10.94 kHz subcarrier frequency spacing, symbol duration of 102,86 ms, frame duration of 5ms [9]. Adaptive modulation and coding is configured on the subscriber stations. Receiver sensitivity is set to -100 dB for both the SS and BS stations.

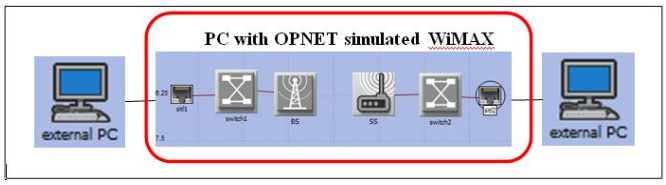


Figure 3. Simulated testbed infrastructure

As shown in Fig. 3, our approach is based on the real-to-real SITL interconnection scenario. Real packets are injected in the simulated network through the SITL interface, flows through the simulated network, are sent outside the simulation through another SITL interface, and are received on a real device. In this way, we can measure the effects of the simulated network on the real packets, obtaining a more accurate evaluation tool for the proposed solutions. The results presented in this paper are obtained by using the following configurations for the simulated network: direct WiMAX links interconnecting the SITL gateways or a double WiMAX link obtained by interconnecting the Base Station with the Subscriber Station through a hybrid mesh node, similar with the relay node used in the real life testbed for the relay based link. As already mentioned, the simulated hybrid mesh node (relay node) was obtained by linking a WiMAX Base Station with a Subscriber Station via Ethernet. The performance of this simulated hybrid mesh node was compared with that of the real hybrid relay node implemented by Thales Company.

IV. TESTS SCENARIOS AND RESULTS

In this section, the test scenarios, used to evaluate the hybrid node performance, and the test results are presented. For these tests, all the real nodes are equipped with smart antennas; while for the simulated nodes both standard and smart antennas were used. The benefits of using smart antennas on WiMAX nodes in static or nomadic configurations are also shown. A first set of experiments was used to evaluate the delay and jitter introduced by the hybrid node, the real one and the simulated one also. A second set of experiments was used to evaluate the performances of the implemented WiMAX network when the smart antennas were used on the WiMAX nodes.

A. *Delay and jitter evaluation for the WiMAX network based on hybrid node*

The tests performed focused on delay and jitter estimation for different segments of the end to end link between Lannion users and UPB users. Round trip time (RTT) value was determined using the ICMP protocol to estimate the delay. The jitter was evaluated using *iperf* application. Standard *iperf* algorithm was used to calculate the jitter. In the experiments presented in this paper only passive 9 beam planar array antennas with sectoral mode were used for smart antennas.

For the RTT and jitter figures, the time in milliseconds is represented on the y-axis, and the packet number on the x-axis. For figures obtained with OPNET simulator, on x-axis the time in minutes is represented.

 ISBN: 978-1-61208-279-0

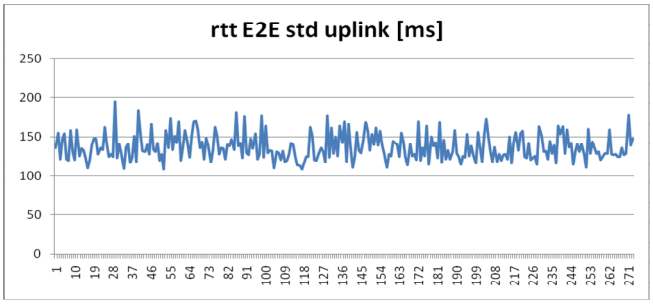


Figure 4. The end to end RTT for standard antennas on the simulated nodes

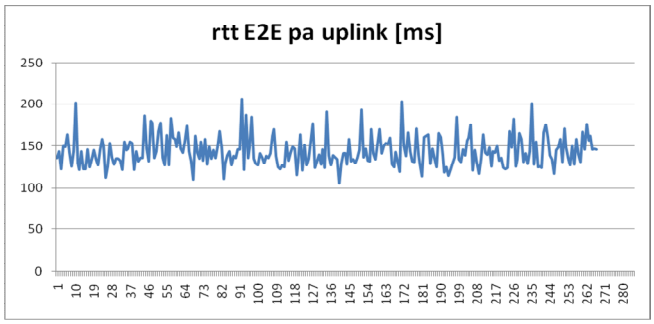


Figure 5. The end to end RTT for plasma antennas on the simulated nodes

The tests were performed for both the standard antennas and smart planar antennas on WiMAX nodes. For the simulated WiMAX link, adaptive modulation and coding was used, and best effort service flows were considered for uplink and downlink.

A first test suite was performed using one WiMAX link in the simulated network and one WiMAX link in the real network. To evaluate the round trip time *ICMP* packets were sent from UPB laptops situated behind the simulated networks. The packets traveled via the simulated WiMAX link, through the VPN tunnel, via the real WiMAX link and then the way back. The RTT obtained for the cases when the simulated nodes were equipped with standard and smart planar antennas were presented in Fig. 4 and Fig. 5. As it was expected the results are similar. One can see that the mean RTT was around 150ms. This value for mean RTT was obtained by taking values for long time intervals.

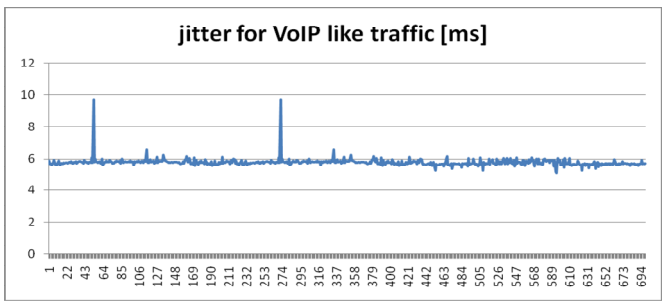


Figure 6. End to end jitter from Lannion to UPB for emulated VoIP traffic

To evaluate the interconnected testbeds behavior for VoIP communications, traffic emulating four VoIP channels was generated using the *iperf* application. The data rate used was of 320 kbps and the packet size was set to 120Bytes. The jitter measured with *iperf* is shown in Fig. 6. The jitter value is around 6 ms which is acceptable for VoIP communications.

The jitter was measured also by sending with *iperf* a 5Mbps UDP data stream from Lannion to UPB. The packet size was set in this case to 512Bytes. In this case the mean jitter was about 5ms.

In order to determine the components of the RTT delay, the individual RTT delays for the links that make up the end to end path were evaluated. The results are presented in Fig.7, Fig. 8, and Fig. 9. The RTT mean delay of the simulated WiMAX link is around 45ms, while the RTT delay for real WiMAX link is around 35 ms. For the simulated WiMAX network the RTT variance is very high because of the variable time it takes for a packet to be translated to the simulation and back. Also, the extra delay for the simulated network is given by the packet conversion time.

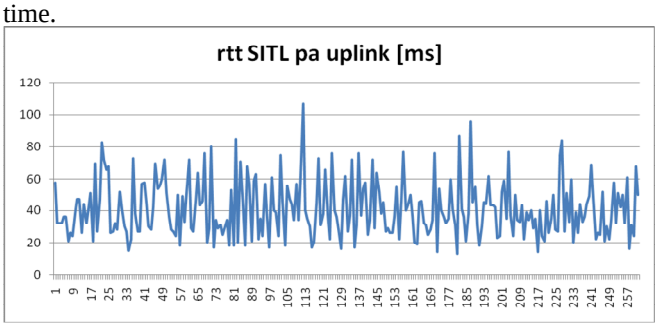


Figure 7. RTT delay for the simulated WiMAX link

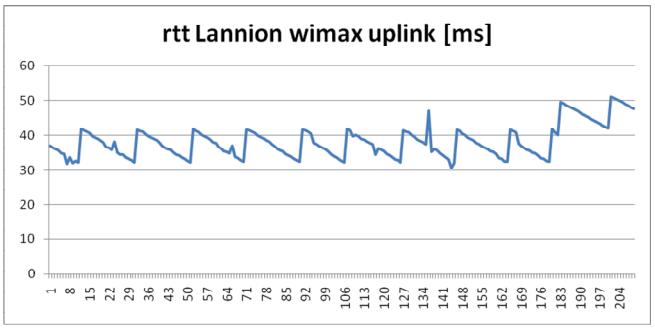


Figure 8. RTT delay for the real WiMAX link

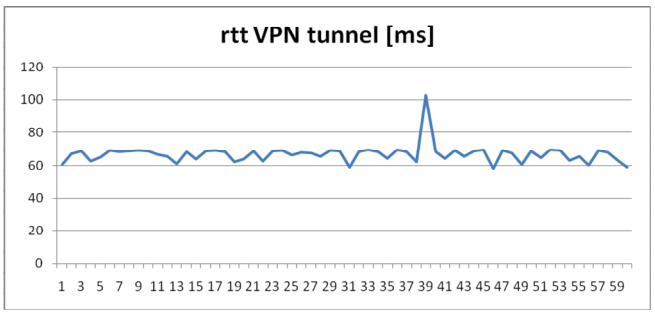


Figure 9. RTT delay for the VPN tunnel

 ISBN: 978-1-61208-279-0

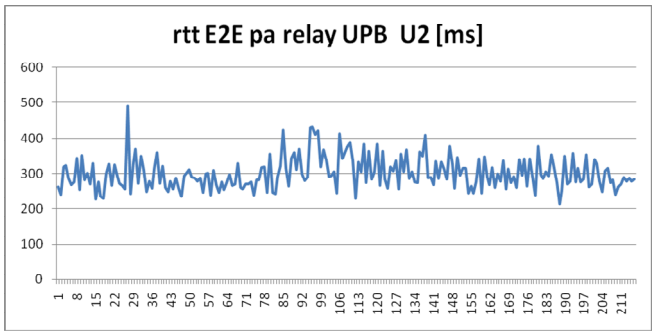


Figure 10. RTT delay for the second test suite

A second test suite was performed using the relay link in the real network and the simulated network. The traffic will travel end to end via two simulated WiMAX links, two real WiMAX links, and via the VPN GEANT tunnel. The RTT delay is shown in the Fig. 10. The mean end to end RTT delay is about 300 ms. This value is obtained by adding the delay introduced by the two additional WiMAX links and by the additional delay introduced by the hybrid simulated node and the relay node.

The jitter obtained in this second scenario is similar with the one obtained in the first scenario. This shows that the jitter introduced by the WIMAX equipments is small compared with the one introduced by the VPN tunnel. This is because the WiMAX link was used at a medium load, 5Mbps, which keeps the jitter at small values.

These measurements showed that the proposed WiMAX hybrid mesh nodes performs well in terms of delay and jitter introduced on the data streams. Such hybrid nodes having routing capabilities can be used to build backhaul infrastructure based on WiMAX technology.

B. Smart antennas benefits evaluation

Using the flexibility of the OPNET simulator, several test scenarios were developed in order to evaluate the gain added by the use of Smart Antennas in such a WiMAX backhaul network. The topology used for these experiments is presented in Fig. 11. The focus was on the simulated section of the network, the obtained results being gathered from the simulations outputs. The simulated part consists of a WiMAX network with one Base Station, BS1, connected to one of the SITL interfaces, one Subscriber Station, SS2, connected to the other SITL interface, and two WiMAX nodes, BS2 and SS1, connected together via Ethernet, which compose the hybrid mesh/relay node. The same WiMAX configuration parameters are used, as the ones in the scenarios presented in section A. The only change is that the same frequency band was used on all the simulated WiMAX nodes.

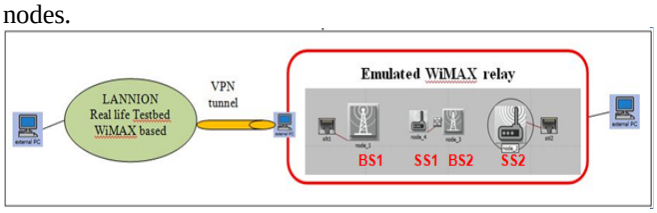


Figure 11. The topology used for smart antenna benefits evaluation

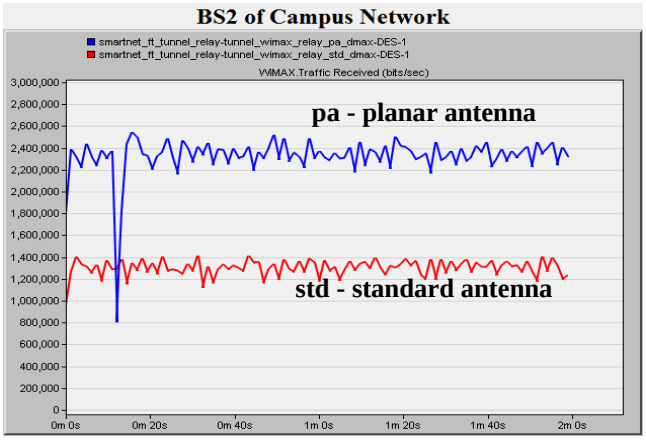


Figure 12. Traffic received by BS2: planar and standard antennas

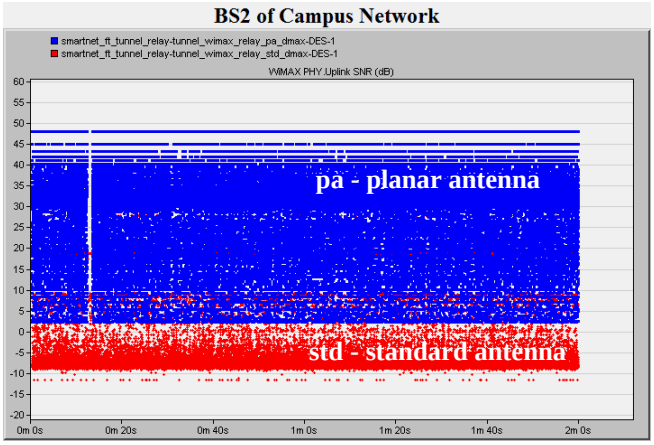


Figure 13. Uplink SINR at BS2: planar and standard antennas

This scenario will illustrate the spatial separation obtained using directional smart antennas. One stream of around 2.5Mbps was sent from the UPB laptop at the Lannion laptop. The traffic will flow through the path SS2-BS2-SS1-BS1 in the simulated network. The simulation was run twice: once with standard antennas on the WiMAX nodes, and once with the WiMAX nodes equipped with smart antennas.

The results are presented in Fig. 12 13 and 14. As it was expected, the results in the case when the directional smart antennas were used on the WiMAX nodes are much better than the ones obtained with standard omnidirectional antennas. This is mainly caused by the fact that the BS1 and SS2 are close to each other and they operate in the same frequency band. For the omnidirectional antennas, the interference level was significantly higher compared with the case when the beam switched smart antennas were used, because of the spatial separation between the BS1 and SS2 provided by the smart antennas. For the real hybrid relay station, due to its timing synchronization, this interference is avoided. The hybrid node used in the simulation, is based only on the spatial separation due to directional smart antennas and to the frequency separation.

 ISBN: 978-1-61208-279-0

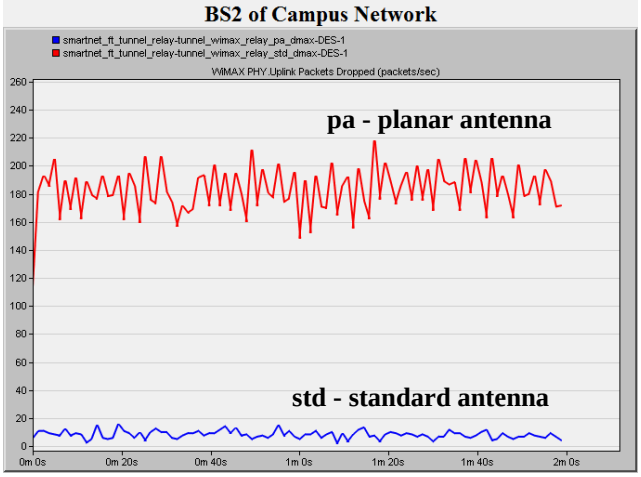


Figure 14. Packets dropped at BS2: planar and standard antennas

In figure 14, the packets dropped at BS2, for the cases when the simulated WiMAX nodes, are equipped omnidirectional standard antennas and directional smart antennas, are shown. The packets dropped by BS2, when omnidirectional antennas are used - red curve, are mainly the consequence of the interference between the nodes composing the hybrid node, BS1 and SS2. The packets dropped by BS2, in the directional smart antenna case- blue curve, are the consequence of the interference generated by the BS1 station and also by the interference generated from the secondary beams of the SS1 directional antennas.

Another advantage of using directional smart antennas is given by the higher gain on the beam direction compared with the equivalent omnidirectional antenna. Other simulation scenarios were used to illustrate this feature. The results showed higher throughput and bigger SINR obtained when the nodes were equipped with smart antennas [9].

V. CONCLUSION AND FUTURE WORKS

This paper presents an evaluation of a mesh network based on WiMAX, with nodes equipped with smart antennas. An experimental platform based on interconnection between an OPNET simulated WiMAX network and a real WiMAX network was built. A hybrid WiMAX mesh node, with routing capabilities, is proposed to be used for building a WiMAX network with mesh topology. An evaluation of the behavior of this hybrid node in a minimal mesh topology is performed. The simulations results presented in this paper illustrate that such an approach can be used for backhaul wireless networks. Also, by using smart antennas for fixed or nomadic WiMAX nodes, a significant performance gain, expressed in terms of capacity and coverage can be obtained for such networks. The proposed WiMAX mesh network with smart antennas is designed only for backhaul network. For mobile users access another radio should be used, or the Base Station should switch to omnidirectional mode to communicate with the mobile users. The smart antenna usage for mobile WiMAX nodes is still under research. The main issue for mobility is to develop a tracking algorithm capable of detecting in real time the best beam to be used to reach the mobile node.

ACKNOWLEDGMENT

This work was supported by the EU FP7 project SMART-Net, no 223937, by the Romanian UEFISCSU PN-2 RU-TE Project no. 18/12.08.2010 and by the EU and Romanian Govern EXCEL project - POSDRU/89/1.5/S/62557.

REFERENCES

[1] IEEE 802.16 Working Group. IEEE 802.16 working group, part 16: Air interface for fixed and mobile broadband wireless access systems|multihop relay specification. IEEE Standard, 2007.

[2] S. Wendt, F. Kharrat-Kammoun, E. Borcoci, B. Selva, A. Tonnerre, and E. Hamadani, “D2.1 - Requirements and Specifications of SMART-Net Target Scenarios", ICT European FP 7 SMART-Net project, May 2010, https://www.ict-smartnet.eu.

[3] S. Wendt, F. Kharrat-Kammoun, E. Borcoci, R. Cacoveanu, R. Lupu, and D. Hayes, “D2.4 - Network Architecture and System Specification", ICT European FP 7 SMART-Net project,Oct. 2010, https://www.ict-smartnet.eu.

[4] C. Cicconetti, I. F. Akyildiz, and L. Lenzini. "Bandwidth balancing in multi-channel IEEE 802.16 wireless mesh networks", InfoCom'07: the 26th IEEE International Conference on Computer Communications, Anchorage, AK, USA, May 6-12 2007, pp. 2108-2116.

[5] D. Ghosh, A. Gupta, and P. Mohapatra. "Scheduling in multihop WiMAX networks", ACM SIGMOBILE Mobile Computing and Communications Review, Volume 12 Issue 2, April 2008, pp. 1-11.

[6] Y. Xu, S. Wan, J. Tang, and R. S. Wol, "Interference aware routing and scheduling in WiMAX mesh networks with smart antennas". SeCon'09: 6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, June 2009.

[7] S. Wan, "Resource Allocation in WiMAX Relay Networks", PhD Thesis, Montana State University, USA, April 2010.

[8] E. Borcoci, M. Constantinescu, S.G. Obreja, A.Baraev, T.Rasheed, and D. E. Meddour. Project Deliverable, “D4.4: System level simulation analyses and performance measures,” ICT FP7 SMART-Net project, May 2011.

[9] S.G. Obreja, I. Olariu, A. Baraev, and E. Borcoci, “Performance Evaluation of a WiMAX Network Using Smart Antennas Through System in the Loop OPNET Simulations” The Eighth Advanced International Conference on Telecommunications, AICT 2012, Sttutgart, Germany, 27May-1June, 2012, pp. 81-86.

[10] J. Mohorko, M. Fras, and Ž. Čučej, ”Real time “system-in-the-loop” simulation of tactical networks,” 16th International Conference on Software, Telecommunicati ons and Computer Networks, SoftCOM 2008, 25-27 Sept. 2008, pp. 105-108.

[11] http://www.plasmaantennas.com/products/selectabeam-portfolio/selectabeam-sc-1000.html [retrieved January, 2013].

[12] S. Wendt, A. Chicot, and M. Skrok “D5.4b Experimental results of real-life Testbed”, ICT European FP7 SMART-Net project Deliverable.

[13] S.G. Obreja, I. Olariu, E. Borcoci, B. Selva, D. Medour, and A. Baraev, “D5.5 - Experimental Results of the Combined Real and Virtual Testbed”, ICT FP7 SMART-Net project, August 2011.

 ISBN: 978-1-61208-279-0

Multi-agents Architecture for Distributed Intrusion Detection

Vinícius da Silva Thiago, Paulo Antonio Leal Rego, José Neuman de Souza
Master and Doctorate in Computer Science
Federal University of Ceará
Fortaleza - CE, Brazil
vsthiago@gmail.com, pauloalr@ufc.br, neuman@ufc.br

***Abstract*—The growing concern about information security in computer networks is responsible for constantly producing new ways to defend them. This work describes the proposal for an Intrusion Detection System architecture that uses agents and an ontology for sharing information. Mobile agents provide a convenient way to distribute the detection process, enabling peer to peer cooperation between network nodes. The ontology provides an organized way of storing and sharing knowledge. To evaluate the proposed solution, the architecture has been implemented using the Java programming language and Java Agent Development Framework.**

Keywords - detection; intrusion; agents

I. INTRODUCTION

The growing number of computer networks applications is responsible for the everyday great diversity and sophistication of attacks and intrusion methods, raising awareness about the safety of these networks. Intrusion detection is one of the key techniques to protect networks and is based on collecting and storing data for auditing systems and networks. According to [1], when detected, an intrusion should be reported to the security manager, and an automatic reply, in order to eliminate the causes and/or the effects of the intrusion, could be triggered.

An Intrusion Detection System (IDS) tries to detect and warn of intrusion attempts to a system or network, in which an intrusion is considered to be an unauthorized or unwanted activity [2]. A centralized IDS runs on a machine in the network in a way that it can collect data from each one of the nodes and then analyze it. However, centralization becomes a major weakness because if the machine crashes, intrusions will not be able to be detected, apart from the fact that the central analyzer can easily become a bottleneck [3].

The distributed detection architectures are more efficient and can solve the problems of centralized architectures. The more sources of information are used to ensure intrusion detection, the more accurate it becomes. The main problem faced by distributed architectures is how to collect and correlate information and then evaluate the security status of the monitored system.

The paper is organized as follows. Section II presents concepts concerning intrusion detection and Section III presents the problem to be treated. Related work is presented in Section IV. Section V describes the proposed architecture, while Section VI details its implementation. Section VII describes the experiment results. In Section VIII, the conclusion and future work are presented.

II. INTRUSION DETECTION

When dealing with intrusion detection, it should be assumed that users and programs activities are observable by auditing mechanisms and that normal activities and intrusions have different behaviors [4]. It is also worth considering that an attacker can try to compromise the IDS itself [5]. Thus, it is important for an IDS to be fault tolerant and/or able to detect problems in its own operation.

In general, the IDSs are composed of four components (sensors, analyzers, database and response units) and are responsible for activities such as monitoring the users and systems activities, auditing systems configuration, accessing data files, recognizing known attacks, identifying odd activities, auditing data manipulation, tagging normal activities, error correction and storing information concerning invaders [6].

Agent systems are composed by a collection of software agents that are autonomous and directed to a goal, located in an organizational context to cooperate through adaptable and flexible interactions and cognitive mechanisms to achieve goals that could not be achieved by a single agent [7]. Mobile agents are defined as processes that can navigate through large networks interacting with machines, gathering information and returning after having carried out the tasks defined by the user [8]. The agents are dynamically updatable, lightweight, have a specific operation and can be used as part of a flexible and dynamically configurable IDS [2].

III. PROBLEM CHARACTERIZATION

An IDS architecture should be simple and effective to provide security against different attacks. According to [9], it is an efficient solution to defend against intrusions cooperatively. It is also important for the IDS to be able to perform its function without compromising the normal operation of the network [10]. The problem then consists in how to build a distributed architecture that is robust to withstand attacks to the very structure of the IDS, enable data sharing between network nodes without creating an excessive overhead in traffic and avoid creating potential bottlenecks.

In the case of a peer-to-peer (P2P) IDS, each host can send detection requests to others without the weakness of a central controller. However, many systems like this only allow hosts on the network to get information from limited sources, such as directly connected neighbors, which can lead to inaccurate decision-making, especially in the case of attacks on multiple hosts [11]. The most important feature of

 ISBN: 978-1-61208-279-0

these types of attacks is that the activity level of the attack in each of the hosts may not be large enough to raise an alarm for the entire network. However, if a distributed IDS can collect and analyze information from multiple hosts, it may be possible to recognize the attack.

One way to implement a distributed IDS is through the use of mobile agents. A host can send mobile agents to others in order to collect relevant information from multiple hosts and to recognize an attack on multiple hosts. According to [12] and [13], the advantages of mobile agent technology include: reducing the overhead of the network, overcoming the problem of delay in the network, executing asynchronously and autonomously, fault tolerance, system scalability and operation in heterogeneous environments.

IV. Related Work

The work presented in [11] showed a proposal for a P2P IDS using mobile agents to achieve a lower processing overhead on the hosts, reduce the risk of a centralized architecture and get more accurate detections. The proposed architecture does not use ontologies and the direct contact between agents is the only way to exchange information about intrusions.

The work presented in [14] proposed the use of semantic techniques in IDS, using ontologies to extract semantic relations between intrusions in a distributed IDS. When IDSagents detect an attack or a suspicious condition, they send messages to the MasterAgent, which can extract the semantic relationships using the ontology and decide whether the activity represents or not an attack. Implemented in a hierarchical architecture, the MasterAgent has shown to be a point of failure, because if an attacker could prevent its operation, intrusions wouldn't be detected. The architecture is efficient in reducing the false positives and false negatives rates and has been implemented using Java Agent Development Framework (JADE).

The work presented in [12] showed the proposal for a distributed IDS using mobile agents and a data mining algorithm to classify network traffic behavior. It proposes the creation of signature detection agents and anomaly detection agents, the latter using data mining techniques. The article also proposed the creation of several classes of agents, in a detection structure similar to [11], with a different technique that classifies network connections according to the level of abnormality found and also proposes that, when detecting new attacks, the signatures are added to the signature detection agents. The architecture does not rely on an ontology. The authors used the JADE framework to implement the proposed IDS.

The work in [15] showed the proposal for a distributed IDS that uses agents and ontology. However, the author had not defined the internal operation of the agents, did not mentioned the use of mobile agents and proposed that the ontology should be left available in a web server, which eventually becomes a single point of system failure. The present paper shows a proposal to define the missing elements, detail the internal operation of the agents and, making some adjustments and changes, to implement an architecture that uses mobile agents very similar to that proposed by the author.

V. Proposed Architecture

This paper proposes a distributed architecture, based on the proposal found in [15] in which agents perform the task of detection by communication and collaboration, using a global ontology. The architecture is organized as a multi-agent detection system which consists of the following classes of agents: sensor, analyzer, manager, ontology, actuator and global ontology. The ontology and manager agents are mobile, while sensor, actuator and analyzer agents are fixed on network hosts and the global ontology agent is a fixed agent which is located on a single host.

The sensor agent captures raw network traffic, transforms it into a pre-defined format and lets it available for the analyzer. This one will analyze the data and apply the detection rules. If an intrusion is confirmed or suspected, two cases are possible: in the first case, a malicious activity is confirmed and the analyzer agent calls an actuator agent to perform the necessary actions; in the second case, the activity is classified as suspicious. In this case, an ontology agent can invoke a sharing of global ontology data by accessing information from the global ontology agent which, if not sufficient for the analyzer agent to decide on suspicions, will make it call a manager agent which will request information related to local suspicious activity from other analyzer agents located in other hosts of the IDS. The operation of the architecture can be seen in Figure 1.

The specification of an ontology separates the data model which defines the intrusion from the operation logic of the IDS, what allows different systems, with distinct operational logics, to share data with no previous agreement on semantics [14].

JADE allows the creation of P2P platforms and the implementation of mobile agents and is under the rules of the Lesser General Public License (LPGL). It is written in Java and offers a large amount of programming abstractions. The structure of the messages exchanged in the communication between agents is based on the Agent Communication Language (ACL) defined by Foundation for Intelligent Physical Agents (FIPA).

VI. Architecture Implementation

In order to observe the operation of the architecture, the proposed agents and ontology have been implemented using the JADE framework. In this section, the implementation of the agents and the ontology are described in details.

A. Sensor Agent Implementation

The sensor agent captures network traffic and saves it in a file in the following format: the Jpcap [16] method Packet.toString( ) is called and the result string is added to the date and time of the reception of the package, each line of the file representing a captured packet. Then it creates an analyzer agent using the JADE command "createNewAgent".

B. Analyzer Agent Implementation

The analyzer agent treats the captured packets one by one and extracts source and destination addresses, source and destination ports (if any), date and time of capturing. It stores the number of packets of a type (with the same features except for date and time) that have been analyzed, the time

 ISBN: 978-1-61208-279-0

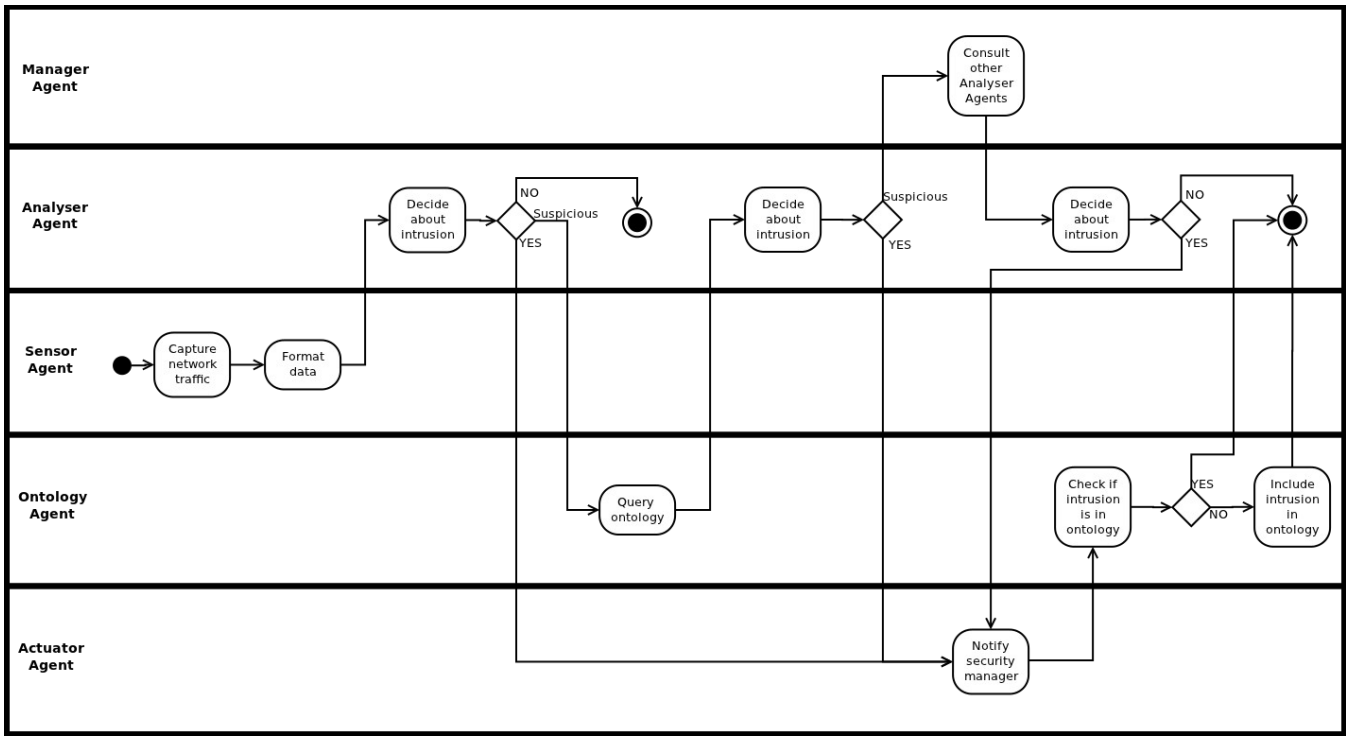


Figure 1. Agents interactions

interval between the last captured packet with these characteristics and if its origin is considered suspicious of attacks.

When implementing the architecture to detect attacks, the following method [15] has been used: for each service, the number of captured packets in the network was counted and two pre-defined limits (L_{min} e L_{max}) were used to decide the traffic's nature. If the observed number of packets was less than L_{min} then it is normal traffic, in this case, the agent temporarily stores the value. Otherwise, if the number exceeds L_{max} then it is a malicious activity. The third case is when the number of packets is greater than L_{min} and less than L_{max} , in which case the traffic is judged as suspect (probably malicious). Consequently, further analyzes should be performed, then it will be necessary collaboration with other agents for more information on the service.

The amount of time that a service has not been accessed until it is noted access in network traffic, has also been considered as a source of information for deciding about suspicious (a simplified solution of the proposed in [17]). The autor ignored in his analysis packages that are not IP protocol and outgoing traffic, in order to reduce the amount of false positives and increase processing speed. However, due to these considerations, some attacks went undetected.

Unlike [17], in the present implementation it has not been disregarded any packet or traffic, to make the system more generic and able to identify more types of attacks. Another difference is that it does not have a training period, executing classification of anomalies based only on the time interval. Thus, if a service goes without being accessed by a time interval greater than a specified value and be accessed, the access is considered suspicious. However, this criterion only classifies as anomalies or normal traffic, not directly identifying intrusions. This identification is performed by the distributed system, by interactions between agents, which, as it uses several sources of information, is expected to reduce the false positive rates.

In case of detecting a number of packets exceeding L_{max}, the analyzer agent creates an actuator agent, passing as arguments the characteristics of the packets that led to the detection. After creating the actuator agent, an ontology agent is created, being passed as arguments its purpose ("inclusion"), which is to include an attack in the ontology, as well as the information passed to create the actuator agent.

In case of detecting a number of packets greater than L_{min} and less than L_{max} , or detecting a time interval of capturing packets of the same type higher than the one set for generating suspicious, the analyzer agent marks the representation of such as suspicious packets and creates an ontology agent, passing as an argument its purpose ("query"), which is to query the global ontology, as well suspicious packets characteristics.

After analyzing each packet, the analyzer agent checks if it has received a message. Upon receiving a message from a native ontology agent stating that it was not possible to confirm a suspicion, it extracts the characteristics of the suspicion and uses them as parameters to create a manager agent. Upon receiving a message from a native ontology agent stating that a suspicion was confirmed, it extracts the characteristics and uses them as parameters to create an actuator agent. Upon receiving a message from a manager agent generated in another host to check the suspects list, it extracts the suspect's characteristics and compares them with its internal information, verifying if the information in the message corresponds to a packet coming from a host it also considers as suspected of generating attacks. If the suspicion

 ISBN: 978-1-61208-279-0

is confirmed, it creates a response message with the content "intrusion occurred". If it is not possible to confirm the suspicion, it creates a response message with the content "not detected". Upon receiving a message from a native manager agent stating that it was not possible to confirm certain suspicion with the others system hosts, it considers that it is not happening an intrusion. Upon receiving a message from a native manager agent stating that a suspicion was confirmed, it extracts the protocol, addresses and ports and creates an actuator agent, passing them as parameters. Then, an ontology agent is created, being passed as arguments its purpose ("inclusion") which is to include an attack in the ontology, as well as the information passed to create the actuator agent.

C. Ontology Agent Implementation

An ontology agent moves to the main container by calling the JADE command "doMove()" and checks the arguments passed in its creation. If the purpose is "query" the agent's goal is to query the global ontology to verify if a suspected intrusion corresponds to one stored in it. If the purpose is "intrusion" the agent's goal is to include the confirmed intrusion in the global ontology. It sends an ACL message to the global ontology agent, whose content is an instance of "suspicion" that consists of two instances of the "host" class (attacker and target) that identify the protocols and network addresses, and two strings that identify the ports numbers. Both hosts and ports match the arguments of the creation of the ontology agent. The message is written in the detection system defined ontology language. Classes "host" and "suspicion" correspond to statements of the global ontology. If the agent's purpose is querying, the message's performative is QUERY_IF, if it is to include an intrusion, the performative is INFORM.

If the message is sent to query, the ontology agent waits for the response message and when receives it, checks if it confirmed the suspicion or not. Then it migrates back to its native host, calling the command doMove( ) and creates an ACL message to the native analyzer agent. The message content is filled with the same parameters used to create the ontology agent, the performative varying to REQUEST if it confirmed an intrusion (because requests that the analyzer agent creates an actuator agent to generate an alarm) or to PROPOSE if it has not confirmed the suspicion (as proposes that the analyzer agent creates an manager agent to check the suspicion in other hosts).

D. Manager Agent Implementation

A manager agent sends a request to get informed about all active containers, creating a list. Thus, it uses the JADE command doMove( ) to move container by container. When migrating to a new host, it sends an ACL message to the local analyzer agent informing the characteristics of the suspicion that was passed as an argument in its creation, asking if it is present in its list of suspects. If the answer to the message is that the suspicion was not confirmed, it migrates to the next host of the list. If the answer is that the suspicion was confirmed, or it has traveled to all hosts on the list, it migrates back to its native container. If the intrusion is confirmed, it creates an ACL message for the analyzer agent, passing information about the protocol, source and destination addresses and ports which were considered as intrusion. Using this information, the local analyzer agent creates an actuator agent to generate the corresponding alarm. If the intrusion is not confirmed, it creates an ACL message to inform the local analyzer agent that it was not possible to confirm the suspicion.

E. Actuator Agent Implementation

The actuator agent extracts the information that has been passed in its creation and adds information regarding the date and time of when the alert is generated, saving this information in a file.

F. Global Ontology Agent Implementation

The global ontology agent is responsible for maintaining the information saved in the ontology and only receives messages written in the ontology language defined for the detection system. Upon receiving a message, it verifies its performative. If it is a message of type QUERY_IF, it is a message asking to perform a query on ontology information. It checks if the attacker's address shown in the query matches an source address in the ontology. After checking, it creates an ACL response message of type INFORM to inform if the detection was confirmed or not. As it only checks the attacker's address when querying the ontology knowledge, the global ontology agent is using the global ontology (in laboratory tests) as a way to represent knowledge about the attackers. If a message received by the global ontology agent is of the type INFORM, its content is a proposed information to be added in the ontology.

G. Detection Ontology Implementation

To define the detection ontology it was necessary to extend the JADE Ontology class. The vocabulary is composed of fifteen strings that represent elements and may be used to represent entities of knowledge that the ontology is intended to describe. The terms that comprise the vocabulary of the ontology are:

- HOST that defines network hosts;
- HOST_ADD that defines a host's address;
- HOST_TYPE_ADD that defines the protocol of the previous term;
- SUSPICION that defines the characteristics of a network stream considered as suspect of being an intrusion;
- SUSPICION_ATTACKER that represents the host suspected of being an attacker;
- SUSPICION_ATTACKER_PORT;
- SUSPICION_TARGET;
- SUSPICION_TARGET_PORT;
- SUSPICION_NUM_PACKETS that represents the amount of packets that were detected accessing a particular service and that generated the suspected intrusion;
- SUSPICION_INTERVAL that represents the time interval that a service has not being accessed;
- INTRUSION that represents an intrusion;
- INTRUSION_TARGET;
- INTRUSION_TARGET_PORT;
- INTRUSION_ATTACKER;
- INTRUSION_ATTACKER_PORT.

 ISBN: 978-1-61208-279-0

After defining the vocabulary, it was necessary to define the schemas that represent the concepts, predicates and agent actions. In the detection ontology, we defined schemas for the concept HOST, the predicate SUSPICION and agent action INTRUSION. Each added scheme is associated with a Java class, so that, when using the defined ontology, expressions indicating the terms need to be instances of these classes.

The concept HOST was associated with the class Host, as well as to primitive schemas HOST_ADD and HOST_TYPE_ADD, both of type BasicOntology.STRING. These associations imply that information passed to the ontology to represent a host must be an instance of Host, that implements the class Concept and has as attributes the strings *Add* and *TypeAdd* and the methods needed to access and assign values to them.

The predicate SUSPICION was associated with class Suspicion, as well as to concept schemas SUSPICION_ ATTACKER and SUSPICION_TARGET, both of type HOST; primitive schemes SUSPICION_ATTACKER_ PORT and SUSPICION_TARGET_ PORT, both of type BasicOntology.STRING; and primitive schemes SUSPICION_NUM_PACKETS and SUSPICION_ INTERVAL, both of type BasicOntology.INTEGER. These associations imply that information passed to the ontology to represent a suspicion must be an instance of Suspicion, which implements the class Predicate and has as attributes the Hosts *Attacker* and *Target*, the strings *AttackerPort* and *TargetPort*, the integers (long) *NumPacotes* and *Interval* as well as methods needed to access and assign values to them.

The agent action INTRUSION was associated to class Intrusion, to the concept schemes INTRUSION_TARGET and INTRUSION_ATTACKER, both of type HOST and to primitive schemas INTRUSION_TARGET_PORT and INTRUSION_ATTACKER_PORT, both of BasicOntology.STRING type. These associations imply that information passed to the ontology representing an intrusion should be an instance of the class Intrusion that implements the class AgentAction and has as attributes Hosts *Target* and *Attacker*, strings *AttackerPort* and *TargetPort* and methods needed to access and assign values to them.

VII. TESTS AND RESULTS

To evaluate the solution, a test lab was prepared. To perform attacks it was used Low Orbit Ion Cannon (LOIC), that performs simple denial of service attacks by sending a sequence of TCP or UDP requests to a target machine. The attacks initiate multiple connections to the same target host and continuously send a predefined string. The group Anonymous used LOIC to carry out attacks on several sites in recent years [18].

The hosts that took part in IDS will be called A, B and C. In the tests, C acted as main container (the container hosting global ontology). Attacks were carried out on the three hosts to see if they could identify the attacks, how the attacks were being detected, and false positives generated by the system. The objective of the tests was to determine whether the proposed architecture could be used in an IDS, although the detection criteria used were quite simple, which do not reflect the reality of the commercial systems currently used. As discussed earlier, the fact that the global ontology agent is located in a specific network node makes this node a weak point of the architecture, however, as seen in [19], JADE allows the main container to be replicated a few times creating redundant containers that take the main-container's place if it becomes unavailable.

There have been performed a total of eight sequences of attacks, in which LOIC was set to flood the target with TCP packets in the first four sequences and UDP packets in the last four. Table 1 summarizes the results.

A. First, Second, Third and Fourth Attack Sequences

In the first sequence, attacks with TCP packets were performed with an amount of packages that exceeded the L_{min} but not exceeded the L_{max} of the analyzer agents. Hosts A, B and C were attacked in that order. At the end of the attacks, the following results were obtained: A did not detect any attack, B detected the attack by means of its manager agents and C detected the attack through its ontology agents. These detections happened according to expectations, since the first host attacked (A) has detected suspicious activity and called its ontology agents, which resulted in no conclusion because there was no information about this attack in the ontology. Thus the manager agent was called, migrated to B and C and also found no information about the attack. Unable to conclude anything about the suspicious, it did not detect the attack. B has detected suspicious activity, called an ontology agent (which also resulted in no useful information) and then called a manager agent that, when migrating to A, received information that the attacker was already considered as a suspect, confirming an attack, calling the actuator agent that generated an alert. After that, the analyzer agent called an ontology agent to add information about the attacker to the global ontology. Host C, when detected suspicious activity, called an ontology agent, that, by consulting the global ontology, confirmed the suspicion of attack and called an actuator agent to generate the alarm.

In the second sequence, attacks were carried out in the same way as at the first, but at the end of the sequence it was performed another attack on A, flooding it with a number of packages exceeding the limit L_{max}. Once again, the results were exactly as expected: B and C generated alarms similar to those of previous sequence and host A has generated two alarms, both due to the detection of packets in excess of the L_{max} limit, an alarm with source on the attacker and an alarm with source on A, which corresponds to the responses of requests from the attacker.

In the third sequence, the attack to A was performed with an amount of packets which exceeded L_{max}, while attacks to B and C were performed with an amount of packets that exceeded L_{min} but did not exceeded L_{max}. At the end of the attacks, A generated two detections when its analyzer agent detected a number of packages exceeding L_{max} coming from the attacker and the generated responses to these requests. B and C detected the attacks through their ontology agents. Once A was attacked first, it has detected the attack, and its actuator agent has called an ontology agent to include the attacker in the global ontology. B and C, when detected suspicious activity, called ontology agents, that by consulting the global ontology, confirmed the suspicions and called actuator agents to generate alarms.

The fourth attack sequence was performed similarly to the first, but in this one it was waited, before initiating the

 ISBN: 978-1-61208-279-0

TABLE I. TESTS RESULTS

	Attack sequence	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Host A	correct detections	0	2	2	2	0	1	1	1
	false positives	0	0	0	12	0	0	0	6
Host B	correct detections	1	1	1	2	1	1	1	1
	false positives	0	0	0	13	0	0	0	9
Host C	correct detections	1	1	1	1	1	1	1	1
	false positives	0	0	0	10	0	0	0	8

attacks, a period of time greater than the interval set in the detection system from which the captured packages are to generate suspicious. Host A generated fourteen alerts: two from the packets coming from the attacker and their responses and twelve false alerts. Host B generated fifteen alerts, two from the packets coming from the attacker and the responses and thirteen false alerts. Host C generated eleven alerts, an alert from the packets coming from the attacker machine and ten false alerts.

B. Fifth, Sixth, Seventh and Eighth Sequences

Sequences five, six, seven and eight were performed in the same way as numbers one, two, three and four, but the LOIC program has been set to flood the targets with UDP packets. Sequence number five obtained the same results as number one. Sequences number six and seven obtained similar results to sequences two and three, but host A generated only one alert, concerning the detection of packets from the attacker machine in an amount that exceeded L_{max}. In the eighth sequence, host A generated seven alerts, an alert for the packets coming from the attacker and six false alerts. Host B generated ten warnings, one concerning the packets coming from the attacker and nine false alerts. Host C generated nine alerts, an alert concerning the packets coming from the attacker and eight false alerts.

VIII. CONCLUSION AND FUTURE WORK

The results of the laboratory tests confirmed that the proposed architecture can be used in intrusion detection processes. All the attacks have been identified by the system, and many have been identified by all hosts attacked. The detection method that considers as decision parameter the time interval the service has not been accessed showed to be able to detect attacks, however, led to the generation of a large number of false positives.

As opportunities for future work, it could be identified: the deployment of a more complex detection, with smarter agents, using statistical anomalies detection identified by managers agents and enabling the creation of attack signatures, which would be stored in the ontology alongside signatures already known; the development of more complex detection ontology, with more parameters to characterize the attacks; the study of the impact of the use of the proposed architecture in network traffic; and the implementation and testing of the architecture with a redundant and fault-tolerant main container.

REFERENCES

[1] R. Puttini, J. Percher, L. Mé, and R. De Sousa, “A fully distributed ids for manet”. In Computers and Communications, Proceedings. ISCC 2004. Ninth International Symposium on, IEEE, Vol. 1, Jun 2004, pp. 331–338, doi: 10.1109/ISCC.2004.1358426.

[2] O. Kachirski and R. Guha, “Effective intrusion detection using multiple sensors in wireless ad hoc networks”. In System Sciences. Proceedings of the 36th Annual Hawaii International Conference on, IEEE, Jan 2003, pp. 57-64, doi: 10.1109/HICSS.2003.1173873.

[3] I. Osman and H. Elshoush, “Alert correlation in collaborative intelligent intrusion detection systems-a survey”. Applied Soft Computing, Vol. 11(7), Oct 2011, pp. 4349-4365, doi: 10.1016/ j.asoc.2010/12/004.

[4] Y. Zhang, W. Lee, and Y. Huang, “Intrusion detection techniques for mobile wireless networks”. Wireless Networks, Vol. 9(5), 2003, pp 545–556, doi: 10.1023/A:1024600519144.

[5] Y. Huang and W. Lee, “A cooperative intrusion detection system for ad hoc networks”. In Proceedings of the 1st ACM workshop on Security of ad hoc and sensor networks, Oct. 2003, pp. 135–147, doi: 10.1145/986858.986877.

[6] D. Farid and M. Rahman, “Anomaly network intrusion detection based on improved self adaptive bayesian algorithm”. Journal of computers, Vol. 5(1), Jan 2010, pp. 23–31, doi: 10.4304/jcp.5.1.23-31

[7] T. Oren and L. Yilmaz, “Synergies of simulation, agents, and systems engineering”. Expert Systems with Applications, Vol. 39(1), 2012, pp. 81–88, doi: 10.1016/j.eswa.2011.06.038.

[8] R. Nakkeeran, T. Albert, and R. Ezumalai, “Agent based efficient anomaly intrusion detection system in adhoc networks”. IACSIT International Journal of Engineering and Technology, Vol. 2(1), Feb 2010, pp. 52-56.

[9] E. Ahmed, K. Samad, and W. Mahmood, “Cluster-based intrusion detection (cbid) architecture for mobile ad hoc networks”. In 5th Conference, AusCERT2006 Gold Coast, Australia, May 2006 Proceedings, <http://eprints.qut.edu.au/33277/> 04.12.13

[10] E. Ferreira, G. Carrijo, R. Oliveira, and N. Araujo, “Intrusion detection system with wavelet and neural artifical network approach for networks computers”. Latin America Transactions, IEEE (Revista IEEE America Latina), Vol. 9(5), Sep 2011, pp. 832–837, doi: 10.1109/TLA.2011.6030997.

[11] D. Ye, Q. Bai, M. Zhang, and Z. Ye, “P2P distributed intrusion detections by using mobile agents”. In Computer and Information Science. ICIS 08. Seventh IEEE/ACIS International Conference on, IEEE, May 2008, pp. 259–265, doi: 10.1109/ICIS.2008.21.

[12] I. Brahmi, S. Yahia, and P. Poncelet, “MAD-IDS: Novel intrusion detection system using mobile agents and data mining approaches”. Intelligence and Security Informatics, Jun 2010, pp. 73–76, doi: 10.1007/978-3-642-13601-6_9.

[13] W. Jansen, “Intrusion detection with mobile agents”. Computer Communications, Vol. 25(15), Sep 2002, pp. 1392–1401, doi: 10.1016/S0140-3664(02)00040-3.

[14] F. Abdoli and M. Kahani, “Ontology-based distributed intrusion detection system”. In Computer Conference, 2009. CSICC 2009. 14th International CSI, Oct. 2009, pp. 65–70, doi: 10.1109/CSICC.2009.5349372.

[15] A. Zaidi, “Recherche et détection des patterns d’attaques dans les réseaux IP à haut débits”. Tesis, Université d’Evry Val d’Essonne, Évry. 2011, 109 f.

[16] K. Fujii, (2007). Jpcap Tutorial. <http://www.eden.rutgers.edu /~muscarim/jpcap/tutorial/index.html>. 04.14.13.

[17] M. Mahoney, “Network traffic anomaly detection based on packet bytes”. In Proceedings of the 2003 ACM symposium on applied computing, Mar 2003, pp. 346–350, doi: 10.1145/952532.952601.

[18] A. Pras, et al. Attacks by ”anonymous” wikileaks proponents not anonymous. Report. University of Twente, Centre for Telematics and Information Technology (CTIT), Dec 2010, <http://doc.utwente.nl/75331/> 04.16.13

[19] F. Bellifemine, G. Caire, T. Trucco, G. Rimassa, and R. Mungenast, Jade administrator’s guide. <http://jade.tilab.com/doc/index.html> 04.12.13.

 ISBN: 978-1-61208-279-0

A Mobile API Solution for Localised Weather Forecast Representation

Paul Dayang and Rebecca Siafaka
Universität Bremen
Dept. of Mathematics / Computer Science
D-28334 Bremen, Germany
{pdayang, rsiafaka}@informatik.uni-bremen.de

Abstract— **In Cameroon, especially in some rural areas, weather forecasting is essential for the everyday activities of the population. Since mobile phones are being extensively used in those areas, the population should be able to acquire the appropriate weather information through such devices. However, most rural populations, such as the Tupuri (Tpuri), use native terms to refer to the weather conditions, mainly based on calendar references. Our paper aims to combine international weather forecast with the local weather description of rural populations in Cameroon, enabling the locals to have fast access such information evolving their everyday activities. We propose an API framework, for mobile development, to extract and present combined and localised weather forecast, showcasing the Tpuri population of Cameroon.**

Keywords-Mobile API; hybrid application development; weather forecast; weather localisation; Cameroon; Tupuri; Tpuri

I. INTRODUCTION

In our era, climate conditions are changing, making the weather forecast an increasingly challenging process. Disasters can be prevented, livestock productivity can be increased, and populations can be prepared for adverse weather conditions, when a forecast is available. Nowadays, especially with the evolution of the Internet, local weather forecasts are widely available through websites that either present related information, or provide with widgets and APIs to retrieve the forecasts from other sources. At the same time, with the development of mobile communications, the users can receive, at any time and anywhere, weather forecasts on their devices.

Considering the popularity the Internet and the mobile communications have gained the last years, we can assume that most of the population worldwide have access to weather forecasts. However, there are populations, especially in rural areas, that have limited access to such information, fact that limits their ability to adjust their activities and protect their lives. One such rural population is the Tupuri (Tpuri), which lives in some areas of Cameroon and Chad. The Tpuri have their own local languages and use their own symbols and rules, to communicate and perform their everyday activities. They also use this language for weather prediction. Although being mostly a rural population, the Cameroonian Tpuri have developed the communication through mobile phones and their mobile network consists a medium that may support the improvement of their everyday activities.

In this paper, we focus on combining the international weather forecasts with the local weather description of the Tpuri population in Cameroon. In order to do so, we suggest an API, for mobile development, which receives local weather forecasts from international resources and presents the information adjusted to the Tpuri native weather reference.

The sections of this paper are divided as follows. In Section 2, we describe the current situation of weather prediction and mobile communications in Cameroon. In Section 3, we give details on the method that the Tpuri people use for weather references. Section 4 includes a description of the selected weather forecast API that we use for our approach. In Section, 5 we present our API solution, giving details on the system architecture and the technology. Finally, in Section 6, we discuss our approach and conclude with our intended future work on the topic.

II. WEATHER PREDICTION AND MOBILE COMMUNICATION IN CAMEROON

Weather forecasts are of a great importance especially for rural populations. However, having access to accurate weather forecasts is not easy for the populations in Cameroon. The *Cercle de Concertation de la Société civile Partenaire du MINFOF/MINEP* (Dialog Group of the Civil Society Partner of Ministry of Forestry and Wildlife, and Ministry of Environment and Protection of Nature) diagnosed the case of meteorological data in Cameroon. The report published in February 2012 [16] stipulates that out of the 58 existing weather stations in Cameroon, only 3 operate. The deteriorated infrastructure is not the only challenge. The lack of technical personnel makes the situation worse. The same report declared that out of the 59 meteorologists in Cameroon (in 2010), 9 went to retirement during 2011 and 15 others will retire between 2012 and 2015.

In Cameroon, the meteorological services are managed by the National Meteorological Service (NMS) of Cameroon, which is under the supervision of the Ministry of Transport. The tasks assigned to this service are, among others, the gathering of climate information, the composition of weather forecasts, the dissemination of meteorological information etc. Except for the air transport, for which the weather data are essential, no more data are available, for

 ISBN: 978-1-61208-279-0

example, for planning the agricultural activities. Indeed, many Cameroonian (e.g. farmers) face the challenge of accessing understandable weather data, because they rely on traditional weather knowledge and, if any, on some forecasts from worldwide weather systems. Providing, as possible, a more accurate weather forecast, requires the collection and analysis of a large amount of data. Hence, the data should be collected over the years in order to generate accurate weather forecasts.

In Cameroon, as in many Sub-Saharan countries, extreme weather and climate events have often various consequences [12], which include loss of lives and livelihoods, damage to infrastructure, increased risk of disease outbreaks, lack of food/water/pasture, mass migration, degradation of the environment, retardation of socio-economic growth, etc.

For example, in August 2012 [9], in the far north region of Cameroon, a sudden flood caused various disasters and victims. Crops were destroyed, cattle disappeared, citizens were forced to leave the flooding areas and people perished under the massive inundations. This disaster might have been prevented, or hindered, if a weather forecaster was available to dispatch information and warn the population in the affected areas. Furthermore, if traditional weather knowledge was used systematically, the population might have been helped to predict and prevent, or at least limit risk related to extreme weather conditions.

If we consider the impact of the livestock sector in Cameroon, the gross domestic product (GDP) in this sector, in 2005, was estimated at 132.8 billion FCFA francs per year (10.36% of GDP) in the primary sector, which corresponds to 2.1% of the national GDP. This sector provides income for about 30% of the rural population (80% of the Cameroonian population lives in rural areas). In 2009, the agricultural sector was estimated for approximately 75.6% of the primary sector with 68.8% for subsistence farming and 6.8% for export crops [17]. As temperature and precipitation are important to agriculture, unexpected weather and climate changes have direct impacts on the livestock productivity.

However, the NMS of Cameroon is poor. It does not provide dynamic data accessible to the users via the Internet, or other medium, like radio or TV, but instead, only static and out of date information on its website. This lack is also followed by the lack of related APIs for data exchange, which could at least enable the acquisition of weather forecast information from other resources.

On the other hand, in Cameroon, the mobile phone market has potential for the development of diversified areas. In fact, a lot of effort has been done to improve the mobile network, making it the most reliable communication network in Cameroon. The situation is the same for many Sub-Saharan countries as well [4]. Mobile operators, which are CAMTEL, MTN and ORANGE, have invested in infrastructural mobile facilities throughout the territory. The mobile network covers over the 80% of the land area. According to the Telecommunications Regulatory Board (TRB) [6], the structure for regulating telecommunications in Cameroon, mobile phone subscribers increased from 0.66% (103279 subscribers) to 44.07% (9,6 million subscribers whereas 6,27% CAMTEL, 50,72% MTN and 43,01% ORANGE) from 2000 to 2010. In July 2012, the first Mobile Virtual Network Operator (MVNO) in Cameroon and also in Sub-Saharan countries, called Set'mobile, has started its activities with an offer for 50,000 subscribers. Thus, we identify a great potential for development in the field. Especially important is the development of tailored applications that aim to support the particular activities of the Cameroonian population and those of other Sub-Saharan countries.

Having considered the current situation of mobile infrastructure and the need for weather prediction in the rural regions of Cameroon, we were further motivated to suggest a model that combines both fields, and gives an important tool for the indigenous populations.

III. NATIVE WEATHER FORECAST TECHNIQUES

Tpuri is one of the, around, 250 tribes existing in Cameroon. They live in the northern part of Cameroon and southern part of Chad, and extend on both sides of the borders between the two countries. The latitude and longitude of the area where the Tpuri live (Tpuriland), in Cameroon, is 10° north and 15° east. The majority of the population lives in rural areas, therefore have agriculture as their main activity to make their living. In this region, the agricultural activity encompasses crops such as millet / sorghum, peanuts, onions, beans and rice. The farming of such crops is highly sensitive to climate changes (drought, flood, etc.). Therefore, weather prediction is of high importance for those populations.

In order to predict the climate changes and give details on the weather conditions, the Tpuri folk rely on indigenous weather knowledge, which is oral and descriptive. In general, the terminology for weather and climate uses words such as dry, wind, rain, humidity, tornadoes etc. Also, the Tpuri rely on such words to describe the weather and climate changes. In order to understand the way this folk describes the weather, we shall consider their calendar description. The Tpuri's calendar is seasonal, therefore, it is based on the local sequence of natural and agricultural events. Referring to the Gregorian calendar, for the Tpuri folk, the new year starts in October. Table I, below, indicates the months in English, as those referred to the Tpuri language [14]. In this table, *few* means "month" or "moon". For example, *few burgi* means "month of dust".

TABLE I. MONTH CORRESPONDENCE OF TPURI AND ENGLISH

few kage	few duugi	few baare	few daa	few darge	few ka'arang
Oct.	Nov.	Dec.	Jan.	Feb.	March

few mene	few burgi	few baa	few yaale	few jon fen sõore wa	few waŋ
April	May	June	July	August	Sept.

Table II provides the meaning of each month, and some related seasonal activities. Furthermore, classified as an event calendar, the table shows the association of natural phenomena, including meteorological events to each month, and the temperatures, which are described orally according to each *few*. The Tpuri only describe the way they feel the weather, i.e. the sensation of the temperature, but not the degrees. They do not use a unit to measure the temperature.

TABLE II. CALENDER OF ACTIVITIES ADAPTED FROM [14]

few	*explanation*	*meteorological events*	*reference*
ancoo	months of cold season, harvest season: peanuts, peas potatoes	cold	October to November
ceere	moon of cool	cold, very cold	December to January
hissi	period of high heat	dry, hot, sun	February to March
burgi	moon of dust	light to heavy wind, dust	April
jo'ge, kabge, mulum	months of sowing, begin for rainy season	light rain	May
baa, yaale	moon of rain / rainy season	rain, humidity	June to July
gumugi, musgware	people are weak and easily catch diseases / outbreak of diseases due to heavy rain	heavy rain, humidity	July to August
hoole gara	red millet harvested at the end of rainy season	rain, humidity	August to September
twale	months when the sun burns	light rain, sun, very hot and humid	September to October

In this table, *ceere*, which means "cold", is the period in which temperatures range between 15° and 20° C, and *hissi*, which means "hot", is the period when temperatures start from 40° C. Consequently, the table builds data records that use, for example, month or temperature, as reference, to match the traditional weather description with those from the existing weather systems. In our API suggested solution, we will use these records, in a database, to describe the weather and climate as the indigenous populations in Cameroon do. But, let us first see what the existing weather APIs provide.

IV. WEATHER FORECAST APIS

Various weather APIs are available online. Mostly, they use the Representational State Transfer (REST), which leverage the HTTP protocol to provide weather data, in JSON or XML format, to other related systems.

In this paper, we use the API provided by World Weather Online [8] (WWO) to showcase our approach. We chose this API only as an example, due to the features it provides, and the detailed weather information for the area of our study, namely the Tpuriland. The provided features are free of charge (for personal and commercial use), therefore suitable for prototyping.

The WWO API provides, for a chosen area, current weather information, as well as for the next 10 days, and the past, up to the 1st July 2008. Some of the information available through this API, is the date/time of the observed weather conditions, the temperature, element description (such as precipitation, humidity, wind speed/direction and atmospheric pressure) and weather description with text and images.

To retrieve information using WWO API, developers shall build simple HTTP requests, including specific keyword variables that refer to specific requested weather, location or data/time attributes. Such attributes are city (filtered by country), town name (filtered by country), latitude and longitude, or IP address. Furthermore, they shall specify the API key (which specifies the licence for its use), the format (XML, JSON or CSV) for the results, and the number of the days for the forecast. The country is an optional value.

An example of an HTTP request, for the region of Kolara in Cameroon, where the Tpuri population lives, looks as follows:

```
http://free.worldweatheronline.com/feed/weather.ashx
?key=xxxxxxxxxxxxxxxx
&q= 10.272019,14.650269
&date=2013-02-14&format=json
```

For the aforementioned query, a portion of the result in JSON format for the 14th February 2013, would look like this:

```
{
  "data": {
    "current_condition": [
      {
        "cloudcover": "0",
        "humidity": "6",
        "observation_time": "12:19 PM",
        "precipMM": "0.0",
        "pressure": "1006",
        "temp_C": "38",
        "temp_F": "100",
        "visibility": "10",
        "weatherCode": "113",
        "weatherDesc": [
          {
            "value": "Sunny"
          }
        ],
        ....
      }
    ],
    "request": [
      {
        "query": "Kolara, Cameroon",
        "type": "City"
      }
    ],
    "weather": [
      {
        "date": "2013-02-14",
        "precipMM": "0.0",
        "tempMaxC": "39",
        "tempMaxF": "102",
        "tempMinC": "24",
        "tempMinF": "74",
        "weatherCode": "113",
                    ....
        ]
    }
}
```

The following figure depicts a screenshot of a weather forecast, requested with the aforementioned query, for the Kolara region, as represented in the website of the API.

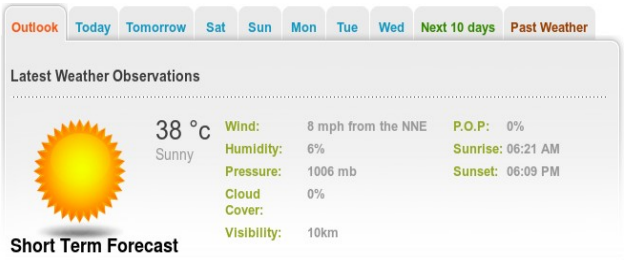


Figure 1. Weather Forecast for Kolara in Tpuriland (14th February 2013)

V. API SOLUTION FOR COMBINED WEATHER FORECAST

Based on the weather forecast that the World Weather Online API provides, we suggest an API model that combines those data with the traditional and indigenous climate knowledge that the Tpuri use, for local weather reference. In our solution, we receive the weather forecast from the selected API, we match the information with the information the Tpuri use to describe the weather conditions, based on the aforementioned calendar, and we then extract combined data.

In the next section, we explain the system architecture for the model. We also give details for the technology to implement the model, as well as the way this can be used later as a mobile application.

A. System Architecture

As already mentioned in Section 3, the selected API offers the forecast data in three formats, namely XML, JSON and CSV. Our model, we choose to explain the approach using the provided JSON format, due to the tools we suggest for the implementation.

The following figure gives an overview of the system architecture.

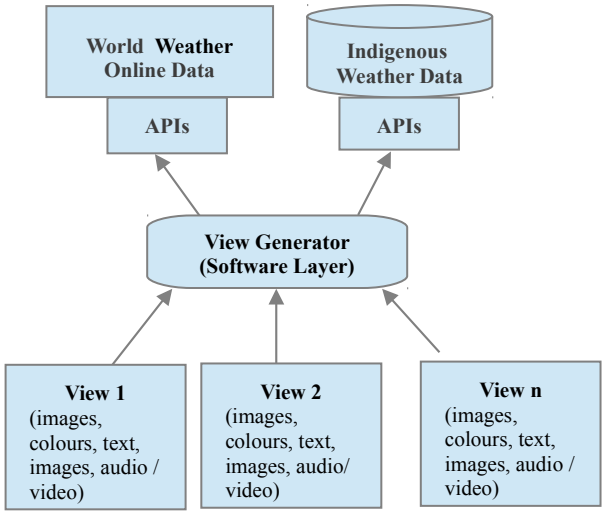


Figure 2. API System Architecture

As the figure shows, the model consists of the following parts:

World Weather Online Data

This part consists of data derived from the World Weather Online API. The data are being retrieved upon demand, for specific date, in JSON format, when Internet connection is available.

Indigenous Weather Data (IWD)

Those data are stored in a database. They reflect, in our case, to the descriptive terms that the Tpuri use for weather reference. Those data are being extracted in JSON format through a related API. The extraction does not require Internet connection.

View Generator

This part is an API that matches the weather forecast data with the Tpuri descriptive data. The view generator receives the JSON input from the aforementioned APIs, and produces a combined JSON file. The weather forecast data include date details that are being matched with the month reference details in the descriptive data. The data process is being performed offline.

Views

The views are representations of the result that the view generator produces. The view generator extracts a JSON file with the combined weather forecast. This file can be then used by a mobile application to print the results in several formats, namely the views.

Since the proposed model uses an Internet connection and a web-based API, it is suitable for a hybrid mobile application. Concerning the bandwidth in the area where the Tpuri live, the download speed goes up to around 0.36 Mbps, and it is sufficient for downloading the appropriate weather information. Once the hybrid app has been installed on the user's mobile device, the typical weather synchronisation process would look like this:

1. The user goes online, when Internet connection is available.
2. Data in JSON format are downloaded from the WWO via the respective API.
3. Indigenous data, in JSON format, are extracted from the IWD via the API.
4. Data from WWO and IWD are combined in JSON output.
5. The JSON format is used from the hybrid app to present the weather forecast in several views with different images, symbols, text, etc..

B. Technology

Here we suggest an implementation approach, describing which technologies may be used to implement an app that makes use of our API framework.

While native apps are implemented in a high programming language (Java or Objective-C etc.), hybrid apps are basically Web applications specifically optimised for use on mobile devices. Hybrid apps provide a good

 ISBN: 978-1-61208-279-0

compromise between native and Web apps to build platform-independent apps. They are Web apps packed as native apps, creating , therefore, a combination of the important features of the native approach and those of the Web.

To implement an app with the aforementioned conceptual approach, we rely on hybrid mobile development using HTML5 and technologies that pertain to it.

Related to the HTML5 technologies, some software developers of the World Wide Web insist that these technologies are revolutionising the Web and its use [1][2]. In fact, HTML5 offers new possibilities to develop Web apps that, although running offline, can process persistent data locally, using Web SQL Database and data in JSON format, through the JQuery library. In order to implement the to implement the conceptual approach, the tool PhoneGap is suitable. PhoneGap [18] is an open source tool that provides a simple and lightweight way for packing Web apps, to operate as native apps, for diverse mobile platforms. It implements a full access to device APIs, such as accelerometer, camera, geolocation, network, alert etc.

This entire approach uses flexible and easy-to-use, web and mobile based technologies, which, by providing weather forecasts, improve the efficient use of Internet through mobile devices, in Sub-Saharan countries like Cameroon, where the Internet connection is still slow and unreliable.

VI. CONCLUSION

In this paper, we provided a model for combining international weather forecast with traditional weather reference information. We consider this model as a threefold. One of the main benefits of this model is that it supports the preservation of the traditional knowledge of Tpuri, which passes orally from one generation to the next, and it is threatened with extinction. The model also allows the integration of both indigenous knowledge and knowledge from well established systems, introducing mobile technologies and supporting the development in this field. Moreover, it provides useful and region-adjusted weather forecast information, assisting the local populations in their everyday activities. The model considers the instability of Internet connection, in most areas in countries like Cameroon (by performing offline, with stored data, when necessary), and exploits the benefits of mobile communications for supporting the locals.

Since most of the Sub-Saharan countries, such as Cameroon, bear similar cultural characteristics, we believe that this solution could be adjusted to other populations, and help to the further development of mobile communications further, for improving the quality of life. Therefore, we intend to continue the research on this field in order to implement, test and adjust our solution to the needs of those populations.

REFERENCES

[1] G. Anthes, "HTML5 leads a Web Revolution" in Communications of the ACM, vol. 55, pp. 16-17, 2012.

[2] M. Pilgrim, HTML5: Up and Running. O'Reilly Media, Inc., 1st ed., Sebastopol CA, 2010.

[3] M. Rao, Mobile Africa Report 2011: Regional Hubs of Excellence and Innovation. Extensia, 2011.

[4] A. Mugoya, African Apps in a Global Marketplace: ideas, observations, tips and some gripes about the African app industry. Asilia, 2011.

[5] Cameroun : 4% de taux de pénétration d'Internet et 9 millions d'abonnés à la téléphonie mobile, http://www.nextafrique.com/sciencestech/technologie/1192-cameroun-4-de-taux-de-penetration-dinternet-et-9-millions-d-abonnes-a-la-telephonie-mobile (January 2013).

[6] Telecommunication Regulatory Board. Informations statistiques sur le marché de la téléphonie (premier et deuxième trimestre 2011), 2011.

[7] Etoo Telecom: C'est finalement ce 21 juillet 2012. http://cameroon-info.net/stories/0,36134,@,eto-o-telecom-c-est-finalement-ce-21-juillet-2012.html (January 2013).

[8] World Weather Online. http://www.worldweatheronline.com (January 2013)

[9] Jeune Afrique. Innondaions meurtrières au Cameroun et au Nigéria. http://www.jeuneafrique.com/Article/ARTJAWEB20120913122017/ (January 2013)

[10] L. Fortnow and R. V. Vohra, "The Complexity of Forecast Testing," ACM SIGecom Exchanges, vol. 7, pp. 1-5, November 2008.

[11] A. Kignaman-Soro, "International Cooperation and Role of THORPEX in the Developing World", the First THORPEX International Science Symposium; Montreal, Canada, December 2004.

[12] INS Cameroun, "Annuaire Statistique du Cameroun 2010," Institut National de la Statistique, 2010.

[13] Kolyang, Parlons Tpuri. L'Harmattan, October 2010.

[14] ACDI Scandale au Ministère des Transports. http://www.acdic.net/ACDIC/component/k2/itemlist/tag/comice%20agropastoral (January 2013).

[15] Direction de la Météorologie Nationale. http://www.meteo-cameroon.net/ (January 2013).

[16] CCSPM La Météo de la météorologie camerounaise : La météorologie au Cameroun est morte. https://groups.google.com/group/sgp-cmr-network/attach/58832513b5bf6241/Communiqu%C3%A9_CCSPM_M%C3%A9t%C3%A9o_Cameroun.pdf?part=4 (January 2013).

[17] Cameroon's National Institute of Statistics. http://www.stat.cm/downloads/CSS/Note_analyse_comptes_nationaux_trimestriels_decembre_2012.pdf (January 2013).

[18] PhoneGap. http://phonegap.com/ (February 2013).

 ISBN: 978-1-61208-279-0

SentiMeter-Br: Facebook and Twitter Analysis Tool to Discover Consumers' Sentiment

Renata Lopes Rosa, Demostenes Zegarra Rodriguez, Graca Bressan
Department of Computer Science and Digital Systems
University of Sao Paulo, SP - Brazil
Email: rrosa@usp.br, demostenes@larc.usp.br, gbressan@larc.usp.br

***Abstract*—Brazilian Consumers' Sentiments are analyzes in a specific domain using a system, SentiMeter-Br. A Portuguese dictionary focused on a specific field of study was built, in which tenses and negative words are treated in a different way to measure the polarity, the strength of positive or negative sentiment, in short texts extracted from Twitter. For the Portuguese dictionary performance validation, the results are compared with the SentiStrength tool and are evaluated by three Specialists in the field of study; each one analyzed 2000 texts captured from Twitter. Comparing the efficiency of the SentiMeter-Br and the SentiStrength against the Specialists' opinion, a Pearson correlation factor of 0.89 and 0.75 was reached, respectively. The polarity of the short texts were also tested through machine learning, with correctly classified instances of 71.79% by Sequential Minimal Optimization algorithm and F-Measure of 0.87 for positive and 0.91 for negative phrases. Another contribution is a Twitter and Facebook search framework that extracts online tweets and Facebook posts, the latter with geographic location, gender and birthdate of the user who posted the comments, and can be accessed by mobile phones.**

***Keywords*—consumer sentiment; Twitter; Facebook; machine learning; social web analysis tool; support vector machines;**

I. INTRODUCTION

Nowadays, people express their sentiments and opinions through social networks and micro-blogs very commonly. There are many sentiment analysis tools for texts posted at micro-blogs, but most of these tools dictionaries are only in English and it is important to consider different people's consumerism vision according to each country and each city.

Analysis tools of emotional texts based on word lists, are best known as ANEW [1], OpinionFinder [2], SentiWordNet, WordNet and SentiStrength [3], of which only the latter analysis tool has support for the Portuguese language and it considers only the unigrams. Sentimeter-Br dictionary covers unigrams (single element/word), bigrams (two adjacent elements) and stopwords (words in a search can be considered irrelevant).

Each field of study requires a specific dictionary, as well as lists of stopwords (e.g., the, of) [4], which need not be analyzed because they do not add value to the performance analysis. Slang and expressions according to each country also need to be considered. It is also important to define the field to be studied,to build a correct dictionary because an only single word may expresses a positive or a negative value or even no kind of emotion.

In [5], an architecture for analyzing in the smartphones field is assembled, analyzing consumers' vision in Twitter, but the study analyzes only one specific micro-blog. Using social networks to analyze sales and features of smartphones or other objects is justifiable because the amount of information is faster to collect, and more data can be gathered.

In [7], a tool captures Twitter data and the polarity of the reviews is analyzed, including slangs that, albeit widely used in social networks, are excluded by different word analyses that have been already mentioned above. A generic dictionary is used in this study. In [8], semantic analysis tools are studied, showing the difficulty to analyze texts from Twitter because there are many slang words and expressions of emotion in the form of symbols. It also shows that some words are not useful to analyze feelings, the so-called stopwords.

The contribution of this work is building a dictionary with the use of regional slangs, emotions, negative words and different verb tenses that have not been considered in other works. A different metric was used, depending on the tense and negative words in the text. The most frequently words were extracted from Google Trends [6] in the last four months to be used in the dictionary.

This work is compared with the SentiStrength tool that has several limitations. The SentiStrength estimates the strength of negative and positive sentiment in short texts. In this work, we joined these two values and turned into a single one.

The polarity of the dictionary was validated by the machine learning technique. The Weka (Waikato Environment for Knowledge Analysis) [9] software was used as a tool for the data analysis.

The algorithms used in Weka were Bayesian networks (Naive Bayes and Bayes Multinomial), Decision trees (C4.5) and Sequential Minimal Optimization (SMO). These algorithms were used to train the data and to decide if a sentence has a positive, negative, neutral or spam value [10].

We built a Twitter and Facebook search frameworks that can be accessed by mobile phones. Thus, these mobile users have access to promotions spread over social networks. The Facebook search framework is complete because it considers the user' s geographical location and their birth dates, if they have configured this information in their Facebook' s user accounts.

Section II provides a theoretical revision of sentiment analysis. Section III deals with the SentiMeter-Br architecture. Section IV presents the machine learning algorithms used in this work. Section V presents the Twitter and Facebook search

 ISBN: 978-1-61208-279-0

framework. Section VI presents the results and discussions and finally, Section VII presents conclusions and the future works.

II. SENTIMENT ANALYSIS

Sentiment analysis, also known as opinion mining, has been studied by researchers, mainly in social webs, such as Twitter. It is a type of computational study of text in natural language, which aims to identify sentiment polarity and intensity of sentiments [11]. The sentiment analysis goal is to classify the polarity of a given text, helping to define if a sentence is positive, negative, or neutral. It is associated to a number in a -5 to +5 scale (most negative to most positive). Each word uses natural language processing or a word dictionary. Another research direction is the subjectivity or objectivity identification [12], but it will not be covered here.

Opinion mining can be used in different topics. A topic in which opinion mining can help is marketing intelligence to know more about people's consuming habits.

Opinion mining in textual data for marketing intelligence can be categorized into three types [13]:

- Early alerting: informs subscribers when a rare, but critical or even fatal condition occurs.
- Buzz tracking: follows trends in topics of discussion and understand what new topics arise.
- Sentiment mining: extracts aggregate measures of positive versus negative sentiment opinion.

This paper analyzes the sentiment mining type and allows learning about buzz tracking, by capturing the words used in tweets.

Sentiment analysis is not a simple task in social networks because the texts can be ambiguous. The use of slangs and ironies is difficult to decipher and to put on a scale as a positive or negative sentiment. So, it is important to have a specific dictionary according to the context because a word can have a negative or a positive value, as in the following texts:

- "Dry hair results from a number of reasons: negative value".
- "The carpet was cleaned and dried: positive value".

A. AFINN word list

There are several word lists to be used in sentiment analysis, with different scales for each word. One of them is the AFINN [14]. Each word in this list has a score from -5 (very negative) to +5 (very positive). Most of the negative words have a minus 2 score, and most of the positive ones have a +2 score. Only the strong obscene words have a -4 or a -5 score, and the entire word list has a bias towards negative words (1598 words corresponding to 65%) [15].

In this paper, a sentiment scale similar to AFINN was used, but with new words in the context to be analyzed, listed by Specialists and captured from Twitter.

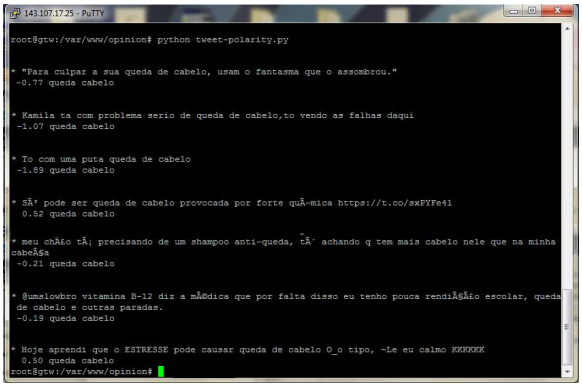


Fig. 1. Sentiment strength value of tweets using the search word *hair loss* = *queda de cabelo* in portuguese.

III. SENTIMETER-BR ARCHITECTURE

Before collecting texts from Twitter, we do a preliminary screening for the most commonly used words in the Internet searches regarding the study area (hair cosmetics) through Google Trends in a four-month period. The dictionary, which is specialized in hair care (shampoos, hair loss, products for greasy hair), began to be formed by AFINN with words of common usage as good, confident, accident. Two specialists were sought to cite most commonly used words concerning about hair cosmetics with a suggestion of values from -5 to +5. These words were added to the dictionary. Their final values were chosen according to an average of the Specialists' suggestion and similar existing words in AFINN list. The most mentioned words by Specialists were adjectives, verbs and some negative words.

Five-hundred texts extracted from Twitter were studied. Some words were also added to the dictionary. The most mentioned by tweets were slangs and some negative words. For other contexts, it will be studied if only five-hundred texts extracted from social networks are necessary or if more texts have to be extracted to make a good classification of polarity.

The Sentimeter-Br dictionary contains 2596 words among which 700 words are tenses, 1600 are adjectives (positive and negative adjectives), 130 are slangs, 116 are emotions and 50 are negatives words (e.g., not, never).

The texts from Twitter that helped to build the dictionary were not used as a test. Three other Specialists validated the dictionary in order not to influence the results. Two thousand more tweets were captured to be classified by the Sentimeter-Br and to have their polarity represented in a numeric value.

The SentiMeter-Br architecture is formed by a script (tweet-polarity.py) in python language to calculate the sentiment strength. The script runs and presents the sentiment strength value as shown in Fig. 1. The texts are extracted from Twitter, by the script, using the Twitter Search API (Application Programming Interface). Data is extracted in JSON (JavaScript Object Notation) [16] format.

It is possible to see the tweets collected by means of a friendly framework through a browser, as is seen in Fig. 2. The results of sentiment strength can be seen in Fig. 1.

The messages crossed the Portuguese dictionary (PT-Br),

 ISBN: 978-1-61208-279-0

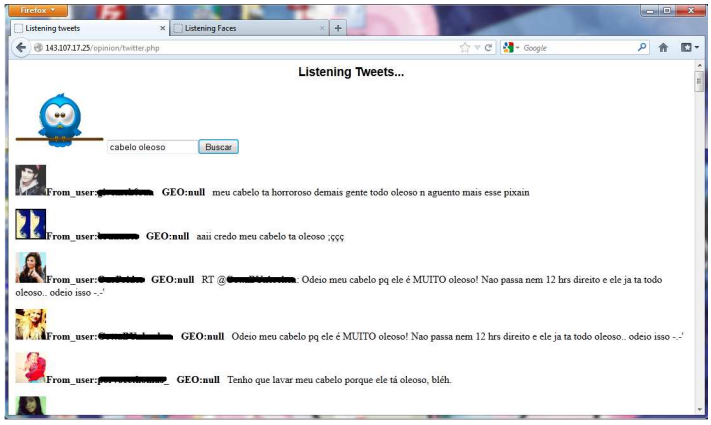


Fig. 2. Friendly framework of collected tweets.

in which each word has a scale from -1 to -5 for negative sentiments and from $+1$ to $+5$ for positive sentiments. It includes emotions with value -1 or $+1$, slang and strong obscene words with values $+5$ or -5. There are separate files for slangs, negative words, negative adjectives, positive adjectives, emotions and tenses (past tense is in a separate file from present tense) in order to facilitate the application of some exceptions, such as the negative rule and the tense rule.

The general sentiment is calculated, $sentiment_{strenght}$, which is the sum of the words divided by the square of the total number of words that are in the PT-Br dictionary, as shown in line 14, in the pseudocode in Table I. The words that are not in the dictionary are considered stopwords, such as words: de (from), para (to), ela (she), among others.

A test was performed with use of unigrams (one word), bigrams (two words) and some trigrams in the dictionary.

The negative words in the tweets were analyzed. If a negative word (contained in NEG-FILE) is followed by a negative adjective word (contained in NEG-ADJ-FILE), as in the example: *not bad*, the word *not* has value $= -1$, the word *bad* has value -3, the result could be $-4/\sqrt[2]{2} = -2,83$, as shown in line 14 of Table I.

However, the words *not bad* should not be so negative because is similar to the word *adequade* that has value = +1. An exception rule of $sentiment_{strenght}$ is implemented, if there are two negative words together and the final value is less than -1 (line 18 of Table I), the lowest value of negative words (bad = -3) is thus added to the final value, multiplied by -1, line 20 of Table I: $-2.83 + 3 = -0.17$.

When it comes to tenses, there is another exception. If a verb is in the past tense (seen in TENSE-FILE, line 10 of Table I), a value of $+1$ is added to the division part (DIV) because verbs in the past tense are less significant in one sentence than a verb in the present tense, as can be seen below:

- my hair looked (0) good $(+3)$ with the shampoo = $3/\sqrt{3} = +1.73$
- my hair looks (0) good $(+3)$ with the shampoo = $3/\sqrt{2} = +2.12$
- I loved $(+3)$ my hair = $3/\sqrt{2} = +2.12$

TABLE I. SENTIMENT STRENGTH PSEUDOCODE

```
1: DIV = 0
2: NEG = 0
3: for i = 1 to N do
4:     read sentiment(word) in ALL-FILES
5:     if (SEARCH word in NEG-FILE) and (SEARCH nextword in NEG-ADJ-FILE) then
6:         LOWER(sentiment(word), sentiment(nextword))
           NEG = NEG + 1
7:         # NEG-FILE = file with words such as NOT, NEVER
8:         # NEG-ADJ-FILE = file with words such as BAD, UGLY
9:     end if
10:    if SEARCH word in TENSE-FILE then
11:        DIV = DIV + 1
12:        # TENSE-FILE = file with LIKED, WAS, WERE
13:    end if
14:    sentiment_strenght = ∑ sentiment / √(len(sentiment + DIV))
15:    # sentiment_strenght: the total of text sentiment value
16:    # sentiment: value of words in the PT-Br dictionary
17:    # len(sentiment): the number of words in the text that are in the PT-Br dictionary
18:    if sentiment_strenght < −1 and NEG > 0 then
19:        for N = 1 to NEG do
20:            sentiment_strenght = sentiment_strenght + (LOWER(sentiment(word), sentiment(nextword))) * −1
21:        end for
22:    end if
23: end for
```

- I love $(+3)$ my hair = $3/\sqrt{1} = +3$
- it was (0) not (-1) good $(+3)$ = $+2/\sqrt{3} = +1.15$
- it is (0) not (-1) good $(+3)$ = $+2/\sqrt{2} = +1.41$

The sentiment strength was measured throughout the dictionary. In the next section, the classification of positive, negative, neutral or spam was performed by the Weka software to assist in results and Specialists' validation.

IV. MACHINE LEARNING ALGORITHMS

Machine Learning is useful to learn patterns through models and templates already scored. This can be used in sentiment analysis, to discover polarity, for example.

In the Weka software, several machine learning algorithms are already integrated and easy to evaluate.

We used Bayesian networks (Naive Bayes and Bayes Multinomial), Decision trees (C4.5) and the Sequential Minimal Optimization (SMO) algorithm to discover if the texts have a positive value, negative, neutral or spam.

Machine learning was used to evaluate the results already obtained from the PT-Br dictionary.

A. Decision Tree

Decision tree is an algorithm that can be used to give the agent the ability to learn and to make decisions.

A decision tree is a model of knowledge in which each branch linking a child node to a parent node is labeled with an attribute value contained in the parent node.

Learning decision trees are examples of inductive learning; they create a hypothesis based on particular instances that generate general conclusions.

The decision trees take as input a situation described by a set of attributes and return a decision that is the value found for the input value.

 ISBN: 978-1-61208-279-0

B. Bayesian Networks

The Bayesian algorithm rating [17] is based on Bayes' theorem of probability. It is also known as Naive Bayes classifier or only as Bayes algorithm.

The algorithm aims to compute the probability of an unknown sample belonging to each of the possible classes.

This kind of prediction is referred to as statistical classification since it is fully based on probabilities.

Multinomial Naive Bayes is a specialized version of Naive Bayes that is designed more for text documents. Whereas simple Naive Bayes would model a document as the presence and absence of particular words, Multinomial Naive Bayes explicitly models the word counts and adjusts the underlying calculations to deal with them.

The distribution is parameterized by vectors $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ for each class y, where n is the number of features (in text classification, the size of the vocabulary) and θ_{yi} is the probability of feature i appearing in a sample belonging to class y.

Parameter θ_y is estimated by a smoothed version of maximum likelihood, i.e. relative frequency counting, as in Equation 1.

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \quad . \tag{1}$$

where:

- $N_{yi} = \sum_{x \in T}$
- $N_y = \sum_{i=1}^{|T|} N_{yi}$ is the total count of all features for class y.

The smoothing prior $\alpha \geq 0$ accounts for features not present in the learning samples and prevents zero probabilities in further computations.

Setting $\alpha = 1$ is called Laplace smoothing, while $\alpha < 1$ is called Lidstone smoothing.

C. Sequential Minimal Optimization

Sequential Minimal Optimization (SMO) is an algorithm described by Platt [18] as using an analytic quadratic programming.

It is an algorithm that solves the Support Vector Machine (SVM) Quadratic Programming (QP) problem without any extra matrix storage and without invoking an iterative numerical routine for each sub-problem.

In [18], the SMO decomposes the overall QP problem into QP sub-problems.

The SMO implements John C. Platt's sequential minimal optimization algorithm for training a support vector classifier using polynomial or RBF kernels.

Multi-class problems are solved using pairwise classification.

Fig. 3. Twitter framework.

Fig. 4. Facebook framework.

V. Message Search Framework on Social Web

A friendly message search framework was used to see the texts extracted from Twitter an Facebook. The users can have access to similar preferences and characteristics by using this framework in order to find promotion products in social web.

An interactive iPhone tool [19] was used to emulate an iPhone to test the frameworks. Fig. 3 shows the Twitter framework and Fig. 4 shows the Facebook framework.

The initial configuration in Facebook framework is necessary because it extracts some data that can only be captured by registered users; these data include geographic location and birthdate. The user needs to enter the Facebook search framework and to inform his/her username and password.

In the case of the Twitter framework, no configuration is required. The Twitter framework was built with the PHP programming language version 5.3 and JSON. It is a simple script and does not use an auto login script as Facebook

 ISBN: 978-1-61208-279-0

Fig. 5. Word cloud of spam words collected in tweets.

because it captures only tweets and Twitter users' id and name.

The Facebook framework was also built with PHP with an auto login script via a client URL Library (cURL) from PHP.

The script uses the FQL (Facebook SQL query) to capture public posts and data user (name, user-id, genre) by Graph API inside the PHP code.

An auto login script was used with cURL, which is a library that lets one make HTTP requests in PHP. It serves to capture Facebook data, such as geographic location and birthdate, which is not public with the use of FQL.

Users can use these frameworks to access promotions posted on Facebook and Twitter and search others members' opinions about a product or a feeling.

VI. Results and Discussions

To validate the work, the results were compared with the SentiStrength tool. The data were evaluated by three Specialists. Each one analyzed two thousand texts captured from Twitter. One hundred tweets were extracted for each word taken from Google Trends.

Comparing the efficiency of the SentiMeter-Br and the SentiStrength against the Specialists' opinion, a Pearson correlation factor of 0.89 and 0.75 was reached, respectively.

Fig. 5 shows the comparison between the SentiStrength tool, SentiMeter-Br and the Specialists (mean) of the 2000 tweets analyzed. Among the texts collected from Twitter, 67% are negative, 21% are positive, 5% neutral and 7% are spam. The spam texts can be useful for companies to analyze their competitors according to the most often cited keywords in tweets, as can be seen in the word cloud of Fig. 6.

Regarding the tweets extracted, 2000 tweets were analyzed by Specialists, 500 tweets were used to help build the Sentimeter-Br dictionary and another 500 were used for training in Weka to help discover the polarity of the 2000 tweets. The Weka classified the sentences as positive, negative, neutral and spam.

The texts (tweets) analyzed underwent string to vector transformation (filter) in Weka software to be able to analyze the words contained in the texts.

The StringToWordVector takes a string and converts it into a vector consisting of the individual words from that string. This is necessary because the Multinomial Naive Bayes classifier, Decision Tree and SMO do not work directly on text, only with separate words.

In Weka results, as seen in Table II, the use of bigrams and trigrams does not much improve the classification as compared with the use of only unigrams in Table III, but the use of stopwords or non-stopwords improved the correct classification in 18.30% with the SMO algorithm as can be seen in Table II. In all the experiments, 10-fold cross-validation was used to evaluate the classification accuracy, and the results of the F-Measure are shown in Table IV.

Spam and neutral texts have a lower F-Measure because on average (of the searched words in Twitter), only 5% are neutral and 7% are spam. The negative texts had a higher F-measure because most of them were negative words.

F-Measure is used to measure the overall performance, combining precision values and the scope of a model in a single formula. The ideal value for average F is 1.

The results with the SMO algorithm presented the best results. Sentimeter-Br can be associated to the classification by the Weka software with SMO as a way of proving the results.

VII. Conclusions and Future Work

Three Specialists conducted the validation, and the results showed the importance of having a specific dictionary according to a context.

The Pearson correlation factor of 0.89 showed the efficiency of the SentiMeter-Br as compared with the SentiStrength. SentiStrength has limitations for the Portuguese language because it does not use a differential calculation for negative phrases, tenses, and specific idioms.

The best results were obtained with the SMO algorithm with higher F-Measure and Correctly Classified Instances. The results from this algorithm proved the polarity of the tweets analyzed.

As future work, we intend to evaluate the SentiMeter-Br in other contexts (business, education, technology, fashion, health).

TABLE II. Percents of Correctly Classified Instances (CC)/ Percents of Incorrectly Classified Instances (CI) of the algorithms with the use of bigrams and trigrams

	Decision Tree	Naive Bayes	Bayes Multinomial	SMO
stopwords	66.66/33.33	64.95/35.04	65.81/34.18	71.79/28.20
no stopwords	63.24/36.75	66.66/33.33	64.10/35.89	60.68/39.31

TABLE III. Percents of Correctly Classified Instances (CC)/ Percents of Incorrectly Classified Instances (CI) of the algorithms with the use of unigrams

	Decision Tree	Naive Bayes	Bayes Multinomial	SMO
stopwords	66.66/33.33	65.81/34.18	60.68/39.31	70.94/29.05
no stopwords	63.24/36.75	66.66/33.33	64.95/35.04	61.53/38.46

TABLE IV. F-Measure of Rating Methods with bigrams/trigrams and stopwords

Algorithm	Positive	Negative	Neutral	SPAM
Decision Tree	0.74	0.88	0.55	0.65
Naive Bayes	0.75	0.85	0.63	0.65
Naive Bayes Multinomial	0.77	0.85	0.69	0.72
SMO	0.87	0.91	0.75	0.79

 ISBN: 978-1-61208-279-0

A way to capture the geographic location in Twitter framework as in the Facebook framework, will be implemented because the public geocode parameter of Twitter, for security reasons, is not shown, unless the user is logged.

In Brazil, Twitter is mostly used by young people. There is difficulty with slangs, repetition of words and, mainly, grammar mistakes. Hence, the studies with Facebook messages will be repeated both in Sentimeter-Br and in Weka.

More texts with spam and neutral classification have to be collected to improve their F-measure.

ACKNOWLEDGMENTS.

The authors thank the University of Sao Paulo for motivating researches in the Computer and Telecommunication Systems area.

REFERENCES

[1] M.M. Bradley, P.J. Lang, Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida (1999).

[2] T. Wilson, J. Wiebe, and P. Hoffmann, Recognizing contextual polarity in phrase- level sentiment analysis. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, Association for Computational Linguistics (2005).

[3] M. Thelwall, K. Buckley, G. Paltoglou, and A. Kappas, Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12) (2010) 2544-2558.

[4] I. A. Braga, Avaliacao da Influencia da Remocao de Stopwords na Abordagem Estatistica de Extracao Automatica de Termos, 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009), So Carlos, SP, Brazil, pp. 18, 2009

[5] W. Chamlertwat, P. Bhattarakosol, and T. Rungkasiri, Discovering Consumer Insight from Twitter via Sentiment Analysis, Journal of Universal Computer Science, vol. 18, no. 8 (2012), 973-992.

[6] Google Trends, http://www.google.com.br/trends, retrieved 18.04.2013.

[7] F. A. Nielsen, A new ANEW: evaluation of a word list for sentiment analysis in microblogs, Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: Big things come in small packages 718 in CEUR Workshop Proceedings 93-98. 2011 May

[8] E. Kouloumpis, T. Wilson, and J. Moore, Twitter Sentiment Analysis: The Good the Bad and the OMG!, In Fifth International AAAI Conference on Weblogs and Social Media (2011).

[9] Weka 3 - Data Mining with Open Source Machine Learning Software,http://www.cs.waikato.ac.nz/ml/weka, retrieved 18.04.2013.

[10] I. Schwab, A. Kobsa, and I. Koychev, Learning User Interests through Positive Examples Using Content Analysis and Collaborative Filtering, Internal Memo, GMD, St. Augustin, 2001.

[11] B. Liu, Opinion Mining and Sentiment Analysis, WEB DATA MINING, Data-Centric Systems and Applications, Part 2, pp. 459-526, 2011.

[12] Bo Pang, L. Lee, Subjectivity Detection and Opinion Identification. Opinion Mining and Sentiment Analysis. Now Publishers Inc, 2008.

[13] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo, Deriving marketing intelligence from online discussion. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD '05). ACM, pp. 419-428, 2007.

[14] F. Nielsen, AFINN-96, Department of Informatics and Mathematical Modelling, Technical University of Denmark (2010).

[15] L.K. Hansen, A. Arvidsson, F. A. Nielsen, E. Colleoni, and M. Etter, Good friends, bad news affect and virality in Twitter. Accepted for The 2011 International Workshop on Social Computing, Network, and Services (SocialComNet 2011), 2011.

[16] JSON (JavaScript Object Notation), http://www.json.org, retrieved 18.04.2013.

[17] M. C. Cirelo, R. Sharoviski, F. Cozman, Coup Gagliardi, and M. H. Coup Veerle. Aprendizado de semi-supervisionado de classificadores bayesianos utilizando testes de independncia. Encontro Nacional de Inteligencia Artificial, Campinas, 2003. SBC 2003 ENIA Anais, Cincia, Tecnologia e Inovao - atalhos para o futuro. Campinas, SBC, 2003. 6 p.

[18] J. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, April 1998.

[19] Interactive iPhone. http://interactiveiphone.com, retrieved 18.04.2013.

 ISBN: 978-1-61208-279-0

A Business Model for Video Transmission Services using Dynamic Adaptation Streaming over HTTP

Demóstenes Zegarra Rodríguez, Renata Lopes Rosa, Graça Bressan
Laboratory of Computer Architecture and Networks - Computer and Digital System Engineering Department
University of São Paulo
São Paulo, Brazil
{demostenes,gbressan}@larc.usp.br , rrosa@usp.br

***Abstract*—A business model for video transmission services is proposed that uses the Dynamic Adaptation Streaming over HTTP (DASH) mechanism using different versions of the same video. These video versions are encoded using different parameters, such as spatial resolution or number of frames per second. The transmission of each video version depends on the capabilities of the network at the end user point. The business model proposed considers different costs for each version of video fragments transmitted. The costs of video resolution upgrade are predefined by the service provider and used during the video streaming. Also, in this model, the user can either accept or reject an upgrade. Subjective tests were performed using different spatial resolutions of the same video, in which the interest level of users to perform a video upgrade is evaluated. Furthermore, the network architecture of the proposed solution is presented and the technical feasibility to deploy the proposed solution in commercial networks is shown. As a consequence, the proposed solution can improve the earnings of the video service provider.**

Keywords-Video Streaming; DASH; Businees Model; Video Quality; QoE; MOS.

I. INTRODUCTION

Nowadays, there are several types of video services applications over IP networks. In this context, video streaming services over the Internet have gained popularity, such as, YouTube [1], Dailymotion [2] among others. This is one of the reasons why the traffic estimated for video on Internet, excluding shared files and games, will reach 61% of the total Internet traffic in 2015 [3].

It is worth noting that, depending on the requirements of video applications and the information to be transmitted, a specific transport protocol can be used. The User Datagram protocol (UDP) is not a connection oriented protocol and therefore the data is not guaranteed to reach its destination; it is generally used for real-time services such as VoIP, some video services or applications of simple request/response e.g., Domain Name System (DNS). The Transmission Control Protocol (TCP) is a connection-oriented mechanism, thereby ensuring that data arrives correctly and in the right order.

Most of the video streaming applications run over HyperText Transfer Protocol (HTTP) which uses TCP as the transport protocol. This is because communication services based on UDP are in some cases intercepted and blocked by firewalls or Network Address Translation (NATs), for this reason, UDP is rarely offered for this type of service.

One problem that arises in the Internet network is congestion. In order to minimize these problems, several congestion control mechanisms for TCP are implemented. Hence, there are many studies aiming to improve its performance, as presented in [4]-[6].

These algorithms can reduce the impairments but none of them solves the problem when TCP detects packet losses. The algorithm reduces the congestion window size and, consequently, the TCP transfer rate decreases. If this new rate is smaller than playback rate, the device' player takes all the buffer information and after that it will enter it into a rebuffering process. During this rebuffering period, no information is displayed and this causes degradation of users' Quality of Experience (QoE).

Different players, such as the open source Adobe Dynamic Streaming [7], Microsoft Smooth Streaming [8], Apple HTTP Live Streaming [9], as well as the players developed by Netflix, and others, use the DASH technique. In adaptive streaming, the video server maintains different versions of the same video, encoded in different bitrates considering spatial and temporal configurations. Also, the video to be transmitted is partitioned into fragments. Then, the client can request different video fragments at different encoding bitrates, considering the network conditions or the users' QoE.

Subjective tests are relevant to quantify the human perception on the quality of voice and video services. As a result, we obtain an index value known as Mean Opinion Score (MOS), which is the mean of the scores granted by at least 15 subjects. Furthermore, providers improve their services based on subjective test results [10].

The most important contribution of this paper is to introduce a new business model for video streaming service using the DASH approach. A test scenario was implemented, in which a video streaming service using DASH was complemented with a billing system. Different resolutions of the same video were stored on the video server for testing purposes. Subjective tests were conducted in a laboratory environment. The goal of these test was to evaluate the users' interest in accepting the suggested video upgrade. Results show that the majority of real users are motivated to improve their QoE watching better spatial resolutions. Additionally, each video version (each resolution) has an identifier (id). For commercial purposes, these identifiers are associated with a predefined cost, and the

 ISBN: 978-1-61208-279-0

upgrade cost is the difference between the cost of two video versions.

In this context, the remainder of this paper is structured as follows. Section II presents the Methods to assess Video Quality considering subjective and objective methodologies. Section III shows the DASH messages and main characteristics. Section IV introduces the Proposed Business Model for video streaming services. Section V presents the results, indicating the users' interest in performing the upgrade and the main functionalities of the proposed solution. Finally, Section VI presents the conclusions.

II. Video Quality Assessment

Video evaluation methods can be classified considering different criteria. The evaluation method can be classified by either using the score given by a subject or the score obtained by an algorithm:

- Objective: these methods use an algorithm or mathematical model to estimate the video impairments. The output of the algorithm could be expressed in different scales. The output of these methods are known as objective metrics.
- Subjective: based on the subject's perception, who grants a video quality score.

The subjective test methodologies and objective metrics are described next.

A. Subjective Test Methodologies

Subjective Methods have been used since the beginning of video quality assessment, and are still valid. They are described in Recommendation ITU-R BT.500-8-11 [11]. This recommendation presents the number of tests and methodologies to correctly conduct the subjective test. The methods are:

- Single Stimulus Continuous Quality Scale (SSCQE);
- Double Stimulus Impairment Scale (DSIS);
- Double Stimulus Continuous Quality Scale (DSCQS);

The ITU-T Recommendation P.910 [12] describes methods of subjective evaluation of video in multimedia applications. The methods presented in this recommendation are:

- Absolute category rating (ACR);
- Absolute category rating with hidden reference (ACRHR);
- Degradation category rating (DCR);
- Pair comparison method (PC).

The description of both, conditions and procedures to produce a video reference is presented in ITU-T Recommendation P.930 [13]. A model to determine the quality of a video transmission using ITU-T P.910 is presented in [14], and its results show how the network variations affect the user's QoE.

B. Objective Metrics

There are different criteria to classify the video quality assessment methods, and they are treated in this section. The ITU defines different categories for the objective methods depending on the type of information considered as the input of the evaluation algorithm, and these categories are the following [15]:

- Media Layer models use either the voice signal or video as an input to estimate the signal quality that determines the end-user QoE. For this method is not necessary to know the network parameters.
- Parametric packet-layer models do not consider the total information content in the packets transmitted. They consider only the headers to estimate the video signal quality
- Planning packet-layer models. In order to perform the estimation quality using this method, prior knowledge of network parameters is required. This method can help the network administrator to optimize the network resources. A proper network planning helps to ensure a good quality of services.
- Bitstream-layer models consider the bit transmitted and, therefore, also use the information used in the parametric packetlayer models.
- Hybrid models are methods that combine two or more of the models described above.

As shown above, there are different objective metrics for video quality assessment, but most of them are not suitable for video streaming over TCP, such as: Peak Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE), Structural Similarity (SSIM) [16], and algorithms based on Region of Interest (RoI) [17] or attentions maps [18, 19]. This reason is because they do not take into account intrinsic characteristics of pauses.

In recent years, few studies have presented metrics based on the application layer parameters, such as the pause intensity metric introduced in [20] or Video Streaming Quality Metric (VsQM) presented in [21]. In this work, we use the VsQM metric in our test scenario.

The VsQM metric proposed in [21] was determined by the parameters: number of pauses, pauses length and weight of the temporal segment during which the pauses occur. These parameters represent the network layer conditions.

The metric $VsQM$ is defined by equation [21]

$$VsQM = \sum_{i=1}^{k} \frac{Wi * Ni * Li}{T} \quad (1)$$

Where:

- N_i is the number of pauses;
- L_i is the average length of pauses, in seconds, that happened in the same temporal segment;
- W_i is a weigh factor which represents the degree of degradation that each segment adds to the total video degradation;
- T_i is the time period in seconds of each segment, and
- k is the number of temporal segments of a video.

III. Video Streaming Service and DASH

Currently, video streaming services do not use any mechanism based on IP networks conditions to improve the QoE.

 ISBN: 978-1-61208-279-0

End user devices are implemented with adaptive video players that use rate adaptation algorithms to minimize the network problems. However, in some cases, this is not enough. In this arena, adaptive streaming plays an important role to improve the user QoE [22]. DASH optimizes and adapts the video characteristics during the video transmission taking into account the network conditions at end user point. Thus, DASH is able to switch from one video fragment to another; the video fragment length is usually between two to ten seconds [23]. Figure 1 shows the messages sent between the client and video server during a DASH process.

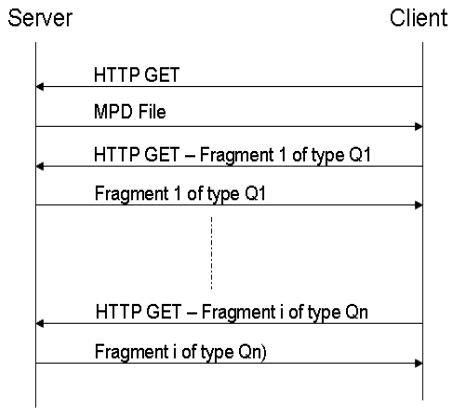


Fig. 1. Messages sent in DASH

The DASH technique presents new challenges and business opportunities for video service providers, telecommunication companies and device developers. Additionally, the research on both video and voice quality metrics are very important, because the adaptive algorithms will choose the video fragment based on these metrics.

The procedure of video streaming over IP network using DASH is described as follows. First, different characteristics of the same video are generated and stored on the video server. These video characteristics can be, for example, the resolution in both spatial and temporal domain, video quality and audio quality. As a second step, each video is logically divided into several fragments; each fragment is identify by a Metadata in the file header. Thus, the client accesses a specific fragment considering its characteristics.

Later, the client sends a HTTP request to the video server ready to attend client requirements. The video server starts the communication sending a DASH-capable video player to the web browser of the client. If an adaptation algorithm is implemented in the player, this algorithm returns a quality metric, which is used to send a new request (standard HTTP GET). Thus, the player requires a video fragment that maps the quality level obtained by the metric. All these steps are performed continuously during the video streaming without any user's action.

IV. The Business Model Proposal

In order to propose a business model for video streaming service, this paper takes into account the DASH technique. The proposal methodology assigns an index value to each version of the same video available in the server. Every video fragment can be identified, and a specific cost can be associated to each fragment.

Figure 2 introduces the solution architecture of the business model proposed for streaming video service using DASH. As depicted in this figure, there are four different versions (A,B,C and D) of the same video. Also, each video fragment is named as *FXi*. Where, *X* represents de video version, and *i* is the fragment number.

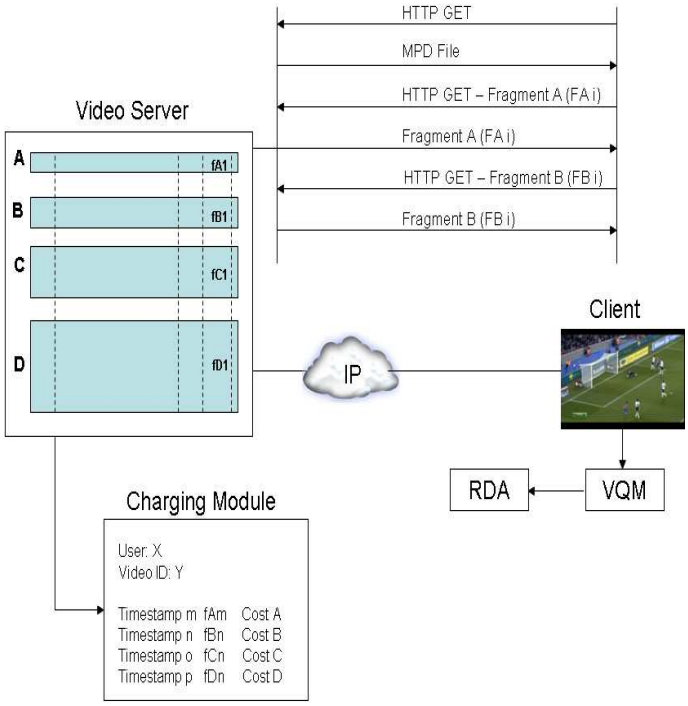


Fig. 2. Architecture of the Business Model for Video Streaming using DASH

From Figure 2, a charging module connected directly with the video server can be observed. In this module, the information sent to the client, regarding all the video fragments is stored and processed. Later, this information needs to be processed in the billing system of the service provider, and finally, charged to the user's account.

The VQM module represents the video quality assessment metric, which is implemented at the application layer, and it is based on the VsQM [21]. In order to estimate the user's QoE, this metric takes in consideration the different buffer status (initial buffering, playing and rebuffering status) in the user's device player. The rebuffering statuses are directly related with the number of pauses and their duration. Then, if the number and duration of pauses increase during a specific period of time, the user's QoE decreases. Further detail about VsQM metric is provided in Section 2.

 ISBN: 978-1-61208-279-0

The module named RDA represents the video Resolution Determination Algorithm that, based on the VQM metric, sends a request for the most appropriated video quality fragment. To perform this action, RDA needs the information contained in the Multimedia Protocol Description (MDP) file.

Some additional considerations about the proposed solution are described below:

- When the RDA module determines that a better video resolution can be received, it sends a message request (HTTP GET - Fragment Ai). The applications running in the video server send the video fragment required (Fragment Ai), and also sends a message with the information regarding the upgrade costs.
- The new video resolution is displayed on the user's device for a time period. This period is programmable, for example, 1 or 3 minutes.
- The user is free to accept or reject the proposed upgrade. Furthermore, there is an option to require not to receive this type of suggestion again.
- If the user accepts the proposed upgrade, the billing module saves the time and the video type transmitted.
- Another input could be used for the RDA module; this new input is related to the networks parameters, such as bandwidth, throughput among others, and is currently under investigation.
- There are other commercial elements that need to be included in the business model, such as content providers and network providers.

V. Test Scenario and Results

In order to determine how the user reacts against a change in video resolution, subjective assessment tests were performed, regarding the user's intention to perform an upgrade. It should be clear that the objective of these tests is not to assess the monetary cost of the upgrade, but the degree of user's interest in performing the video resolution upgrade.

Considering the architecture of Figure 2, tests were performed using only three categories of video resolution, which were previously stored in the video server. These categories are resolution A (A), resolution B (B) and resolution C (C), where A is the lowest resolution, B is the intermediate resolution and C is the highest resolution.

We used the following video quality upgrade scenarios: from A to B, B to C, and A to C. Each upgrade scenario was evaluated using a score into a four-point scale, in which score one represents that user has no interest. Score four represents the case in which the user is very interested in performing the upgrade.

The scale used in the subjective test is presented in Table I.

Carrying out subjective tests involved 23 volunteers who are economically independent, for example, they do not need a third person's authorization to make a payment. Additionally, none of them have any sight problem.

It is considered that the new resolution video (for example, resolution B, in the case of an upgrade of type A - B) is shown for a period of time, i.e., one minute, and in parallel, the upgrade cost is sent to the user screen. If the user accepts the upgrade cost, the charging module of Figure 2 starts recording the new resolution.

TABLE I
Scale used in Subjective Tests

Grading Value	Interest on video resolution upgrade
1	I am sure I have no interest
2	Not right now, maybe later
3	I have interest
4	I am very interested

The global subjective test results of video resolution upgrade intention are presented in Table II.

TABLE II
GLOBAL RESULTS OF SUBJECTIVE TESTS ON VIDEO RESOLUTION UPGRADE

Upgrade Scenario	Score (Mean)
A - B	2.91
B - C	2.13
A - C	3.30

Table III presents the subjective test results in more detail, indicating the number of votes for each option.

TABLE III
RESULTS OF SUBJECTIVE TESTS ON VIDEO RESOLUTION UPGRADE

Scale	Upgrade A-B (number of votes)	Upgrade B-C (number of votes)	Upgrade A-C (number of votes)
1	3	7	2
2	5	9	5
3	6	4	8
4	9	3	10

It can be concluded from the results that users have more interest in performing the upgrade; when they are watching the smallest resolution and the possibility to watch the highest resolution appears. Thus, for example, upgrade scenario "A to C" reached the highest score. Also, Table II allows observing that in all upgrade cases, the mean score is higher than 2; this fact indicates that many users intend to accept the upgrade.

VI. Conclusions and Future Work

The DASH technique permits deploy different solutions, such as the business model proposed for video streaming services. The network architecture of the solution proposed only adds the charging module to the DASH architecture. For this reason, the deployment in commercial networks over Internet or cellular networks is feasible. The implementation of the solution does not represent great costs, and the technical deployment is not complicated.

In the solution proposed, the user is free to accept or reject the video resolution upgrade offered. Thus, the user has the control and takes the decision.

 ISBN: 978-1-61208-279-0

Results of subjective tests show that real users are interested in performing an upgrade, and they can accept to pay a bit more for the spatial resolution upgrades. Thus, video streaming service providers using the business model proposed can increase their income due to the different video rates.

This work shows that video quality metrics play an important role in video streaming services, because their results can be used in different solutions in order to improve the user's QoE.

As future work, conducting subjective tests using more spatial resolution of the same video as test material are proposed, and also different video contents. Also, videos with different temporal resolutions will be included in the test material. Additionally, to have statistically valid results, the number of volunteers to perform the test will be increased.

ACKNOWLEDGMENTS

The authors thank the Laboratory of Computer Architecture and Networks (LARC) at University of São Paulo for the motivation to research in the area of Computer and Telecommunication Systems. This work was supported by FAPESP (Foundation for Researching Support of São Paulo -Brazil). FAPESP project number: 2011/12724-8

REFERENCES

[1] *Youtube, available at http://www.youtube.com*. Retrieved Apr. 15, 2013.
[2] *Dailymotion, available at http://www.dailymotion.com/*. Retrieved Apr. 15, 2013.
[3] *Cisco System, Visual Networking Index*. White Paper. Jun. 2011.
[4] O. Hiroki, H. Hisamatsu and H. Noborio, *Design and Evaluation of Hybrid Congestion Control Mechanism for Video Streaming*. 11th IEEE International Conference on Computer and Information Technology, 2011.
[5] H. Hisamatsu, G. Hasegawa, and M. Murata, *Non bandwidth-intrusive video streaming over TCP*. Proceedings of 2011 Eighth International Conference on Information Technology, pp. 78-83, Apr. 2011.
[6] L. Cai, X. Shen, J. Pan, and J.Mark, *Performance analysis of TCP friendly AIMD algorithms for multimedia applications*. IEEE/ACM Transactions on Networking, pp. 339-355, Apr. 2005.
[7] *Adobe, HTTP Dynamic Streaming on the Adobe Flash Platform, http://www.adobe.com/products/httpdynamicstreaming*. Retrieved Apr. 18, 2013.
[8] *Microsoft, Smooth Streaming technical overview*.
[9] *Apple, HTTP Live Streaming., http://developer.apple.com/resources/http-streaming*. Retrieved Apr. 18, 2013.
[10] H-J. Park and D-H. Har, *Subjective Image Quality Assessment based on Objective Image Quality Measurement Factors*, IEEE Trans. Consumer Electron., Vol. 57, no. 3, pp. 1176-1184, Aug. 2011.
[11] ITU-R Recommendation BT.500-11, *Methodology for the Subjective Assessment of the Quality of Television Pictures*.
[12] ITU-T Recommendation-P.910, *Subjective Video Quality Assessment Methods for Multimedia Applications*, Geneva, Sep. 1999.
[13] ITU-T Recommendation-P.930, *Principles of a reference impairment system for video*, Geneva, Sep. 1996.
[14] ITU-T Recommendation-P.931, *Multimedia communications delay, synchronization and frame rate measurement*, Geneva, Nov. 1998.
[15] A. Takahashi, D. Hands, and V. Barriac, *Standardization activities in the ITU for a QoE assessment of IPTV*. IEEE Commun. Mag., Vol. 46, no. 2, pp. 78, Feb. 2008.
[16] A. Wang, and C. Bovik, and H. Sheikh, *Image Quality Assessment: From Error Visibility to Structural Similarity*. IEEE Transactions on Image Processing, Vol. 13, no. 4, pp. 600-612, Apr. 2004.
[17] H. Kwon, H. Han, S. Lee, W. Choi, and B. Kang, *New Video Enhancement Preprocessor Using the Region-Of-Interest for the Videoconferencing*. IEEE Trans. Consumer Electron., Vol. 56, no. 4, pp. 2644-2651, Nov. 2010.
[18] J. You, A. Perkis, M. Gabbouj, and M. M. Hannuksela, *Perceptual quality assessment based on visual attention analysis*, in Proc. ACM Int. Conf. Multimedia, Beijing, China, pp. 561-564, May 2009.
[19] A. K. Noorthy and A. C. Bovik, *Visual importance pooling for image quality assessment*, IEEE J. Select. Topics Signal Processing, Vol. 3, no. 2, pp. 193-201, Apr. 2009.
[20] T. Porter and X. H. Peng, *An Objective Approach to Measuring Video Playback Quality in Loss Networks using TCP*, IEEE Communications Letters, Vol. 15, no. 1, pp. 76 - 78, Jan. 2011.
[21] D. Rodríguez, J. Abrahão, D. Begazo, R. Rosa, and G. Bressan, *Quality Metric to Assess Video Streaming Service over TCP considering Temporal Location of Pauses*, IEEE Transaction on Consumer Electronic, Vol. 58, Issue: 3, pp. 985 - 992, Aug. 2012.
[22] O. Oyman and S. Singh, *Quality of experience for HTTP adaptive streaming service*, IEEE Commun. Mag., Vol. 50, Issue:4, pp. 20 - 27, Apr. 2012.
[23] A. Begen, T. Akgul, and M. Baugher, *Watching video over the Web: Part 1: Streaming Protocols*, IEEE Internet Computer, Vol. 15, Issue:2, pp. 54-63, Mar. 2011.

 ISBN: 978-1-61208-279-0

On the Capacity of a Cognitive User with Subcarrier Collisions over Rayleigh Fading Channels

S. Ekin, E. Serpedin
Dept. of Electrical and Computer Eng.
Texas A&M University
College Station, Texas, USA
email: sabitekin@gmail.com,
serpedin@ece.tamu.edu

M. Abdallah, K. Qaraqe
Dept. of Electrical and Computer Eng.
Texas A&M University at Qatar
Doha, Qatar
email: mohamed.abdallah@qatar.tamu.edu,
khalid.qaraqe@qatar.tamu.edu

***Abstract*—The paper evaluates the capacity of a cognitive user in an orthogonal frequency-division multiplexing-based spectrum sharing communication system that assumes random subcarrier allocation and absence of spectrum sensing information at the secondary (cognitive) user. In the absence of the primary user's channel occupation information, i.e., no spectrum sensing mechanism is used, the secondary user randomly accesses the subcarriers of the primary network and collides with the primary users' subcarriers. The capacity of a secondary user under such a random access that assumes subcarrier collisions is evaluated herein paper and used as a performance benchmark to investigate the proposed communication scheme over the Rayleigh fading channel.**

Keywords-cognitive radio; OFDM; spectrum sharing; capacity; random access;

I. INTRODUCTION

Recent measurements have highlighted that the radio frequency (RF) spectrum is being under-utilized. Therefore, the cognitive radio (CR) technology has recently received huge interest because it has the potential to yield more efficient RF spectrum utilization [1]. This paper focuses on evaluating the capacity of a cognitive user in an orthogonal frequency-division multiplexing (OFDM)-based underlay spectrum sharing communication system, where primary users (PUs) are allocated a higher priority to access the RF spectrum than secondary users (SUs), and the coexistence of primary and secondary users is enabled under the PU's predefined interference constraint [2], also called interference temperature.

One of the critical issues faced in the adoption of CR networks is to know whether at a certain physical location and moment of time the RF spectrum is occupied by PU(s), i.e., if there is a sensing mechanism in place for the available spectrum [3], [4]. The challenge in deploying spectrum sensing mechanisms is due to the uncertainties ranging from channel randomness at device and network-level, hidden primary users, and issues pertaining to sensing duration and security [5].

Because of the difficulties faced in the acquisition of the spectrum sensing information, this paper focuses on assessing the capacity of a cognitive user in an OFDM-based CR spectrum sharing communication system that assumes random allocation and absence of the PU's channel occupation information, i.e., lack of spectrum sensing information. In other words, the SU's subcarriers are allowed to collide with PUs' subcarriers. As a major benefit of the proposed random access scheme, random subcarrier utilization helps to uniformly distribute the SU's interference among the PUs' subcarriers, and hence, to equalize uniformly the performance of all cognitive users across the whole network. So far, no studies have been reported to assess the effects of subcarrier collisions in OFDM-based cognitive spectrum sharing systems. Therefore, there is a critical need for a more comprehensive system analysis including the development of a stochastic model to capture the subcarrier collisions and protection of the operation of PUs in OFDM-based cognitive systems.

The outcomes of the analysis conducted herein paper will help in understanding the performance limits of random access OFDM-based cognitive communication systems and could be also utilized as performance benchmarks to assess the performance of other cognitive spectrum sharing systems that assume the availability of spectrum sensing information. The rest of this paper is organized as follows. Section II introduces the notations and assumptions that will be used throughout this paper, and describes the basic features of the adopted OFDM-based cognitive radio system. Section III is dedicated to the evaluation of the capacity of a secondary cognitive user in the presence of subcarrier collisions and Rayleigh flat fading channels. Computer simulations corroborating the proposed capacity evaluation study are presented in Section IV. Finally, Section V concludes the paper with some future possible extensions of the current results.

II. SYSTEM MODEL

In the considered system model, the primary and cognitive (secondary) networks consist of N PUs with a primary base station (PBS), and M SUs with a secondary base station (SBS), respectively. To preserve the quality of service (QoS) requirements of PUs in the proposed random access OFDM spectrum sharing communication system, the interference

 ISBN: 978-1-61208-279-0

power levels caused by the SU-transmitters at the primary receiver (PBS) are enforced to be smaller than a predefined interference temperature (IT) (Ψ_i) at the i-th subcarrier.

The channel power gains from the mth SU to SBS and PBS are represented in terms of variables h_m and h_{mp}, respectively. Also, g_n and g_{ns} denote the channel power gains from the nth PU to PBS and SBS, respectively. All the channel gains are assumed to have unit mean and be independent and identically distributed (i.i.d.) flat Rayleigh fading random variables. It is also supposed that perfect information about the interference channel power gains, h_{mp}, is available at SUs. The SU can access the channel side information (CSI), through various means such as the channel reciprocity condition [6], [7], mediate band mechanism or cognitive radio network manager that coordinates the operation of PBS and SU [2]. The thermal additive white Gaussian noise (AWGN) at PUs and SUs is assumed circularly symmetric complex Gaussian stochastic process with zero mean and variance η, i.e., $\mathcal{CN}(0,\eta)$. The parameters $h_{m,i}$, $h_{mp,i}$, $g_{n,i}$ and $g_{ns,i}$ will represent the channel power gains associated with the ith subcarrier.

The total number of available subcarriers in the primary network is denoted by F. The subcarrier set of each PU is assigned by ensuring the orthogonality among the sets of subcarriers for all PUs, F_n^P for $n = 1, \ldots, N$. SU randomly accesses the subcarriers from the available subcarriers set F without knowing which channels (subcarriers) are occupied by PUs. Therefore, SU will collide with the subcarriers of the PUs with a certain probability. The probabilistic model adopted for the number of subcarrier collisions is very general and follows a multivariate hypergeometric distribution. During the evaluation study of the SU capacity (to be described in Section III), it will be assumed that there is only a single SU (which can be any SU in the system) in the cognitive network. The reason for focusing on such a simplified set-up is because such a framework enables to conduct a more simplified analysis without introducing a very cumbersome and hard-to-follow analysis, and to draw very general conclusions. In the same time, the proposed framework can be easily extended to multiple SUs with the assumption of no mutual interference among SUs.

III. Cognitive User Capacity Analysis

A. Secondary User Capacity with Subcarrier Collisions

The Rayleigh fading channel model is adopted to investigate the impact on performance of the system parameters and to evaluate the expressions for the probability density function (PDF) and cumulative distribution function (CDF) of SU capacity. We will focus first on the SU capacity expression with subcarrier collisions.

1) Probability Mass Function of the Number of Subcarrier Collisions: If any arbitrary (consider the mth) SU randomly accesses F_m^S subcarriers from a set of F available subcarriers without replacement while $\sum_{n=1}^{N} F_n^P$ subcarriers are being used by the N PUs, then the joint probability mass function (PMF) of the number of subcarrier collisions, $\mathbf{k}_m$, follows the modified multivariate hypergeometric distribution:

$$p(\mathbf{k}_m) = \binom{F_f}{k_{fm}} \binom{F}{F_m^S}^{-1} \prod_{n=1}^{N} \binom{F_n^P}{k_{nm}}, \tag{1}$$

where the notation $\binom{\cdot}{\cdot}$ stands for the binomial coefficient, and $\mathbf{k}_m = [k_{1m}, k_{2m}, \ldots, k_{Nm}, k_{fm}]^T \in \mathbb{Z}_{0+}^{N+1}$ represents the number of collisions of the mth SU with N PUs and with the collision-free subcarriers, k_{fm}. Parameter $F_f = F - \sum_{n=1}^{N} F_n^P$ stands for the number of free subcarriers.

2) Capacity with Collisions: Let k_{nm} denote the number of (mth) SU's subcarriers that collide with the nth PU's subcarriers, then the capacity of SU with subcarrier collisions can be expressed as:

$$C_m = \sum_{i=1}^{k_{1m}} \log\left(1 + S_{m,i}^{I,1}\right) + \cdots + \sum_{i=1}^{k_{Nm}} \log\left(1 + S_{m,i}^{I,N}\right) + \sum_{i=1}^{k_{fm}} \log\left(1 + S_{m,i}^{NI}\right) = \underbrace{\sum_{n=1}^{N} \overbrace{\sum_{i=1}^{k_{nm}} C_{m,i}^{I,n}}^{C_m^{I,n}}}_{C_m^I} + \underbrace{\sum_{i=1}^{k_{fm}} C_{m,i}^{NI}}_{C_m^{NI}} \tag{2}$$

where $S_{m,i}^{I,n}$ and $S_{m,i}^{NI}$ represent the signal-to-interference plus noise ratio (SINR) for the ith subcarrier of the mth SU with "interference" and "no-interference" from the nth PU, respectively. We make the remark that $S_{m,i}^{NI}$ is indeed the signal-to-noise ratio (SNR) for the ith subcarrier. However, throughout this paper to emphasize the subcarrier collision and collision-free cases, it will be referred to as the SINR with "no-interference" from PU. All logarithms herein paper are with respect to Euler's constant e.

B. Cognitive Capacity over Rayleigh Fading Channels

The peak power interference constraint is considered herein paper, and an adaptive scheme is used to adjust the transmit power of SU to preserve the QoS of PUs. Hence, the transmit power of the mth SU corresponding to the ith subcarrier is given by $P_{m,i}^T = \min\{P_{m,i}, \Psi_i / h_{mp,i}\}$, for $i = 1, \ldots, F$.

Define the variable $\lambda_{m,i} := h_{m,i} P_{m,i}^T$, then the received SINR of the mth SU's ith subcarrier takes the form:

$$S_{m,i}^{I,n} = \frac{\lambda_{m,i}}{I_{n,i}^P + \eta}, \quad \text{for } n = 1, \ldots, N, \tag{3}$$

where $I_{n,i}^P = P_{n,i} g_{ns,i}$ denotes the mutual interference caused by nth PU on the ith subcarrier. In (3), $S_{m,i}^{I,n}$ represents the SINR when subcarrier collision happens. Therefore, when there is no collision, i.e., the subcarrier is

not being used by two users, there is no interference caused by PUs. Thus, $S_{m,i}^{NI} = \lambda_{m,i}/\eta$.

The PDF and CDF of $S_{m,i}^{NI}$ are given, respectively, by

$$f_{S_{m,i}^{NI}}(x) = \frac{\eta e^{-\frac{\eta x}{P_{m,i}}}}{P_{m,i}} \left[1 - e^{-\frac{\Psi_i}{P_{m,i}}} \left(\frac{(\eta x)^2 + \Psi_i \eta x - \Psi_i P_{m,i}}{(\Psi_i + \eta x)^2}\right)\right],$$

$$F_{S_{m,i}^{NI}}(x) = 1 - e^{-\frac{\eta x}{P_{m,i}}} + \frac{\eta x}{\Psi_i + \eta x} e^{-\frac{\eta x + \Psi_i}{P_{m,i}}}.$$

The derivations of the expressions for the PDF and CDF are omitted due to lack of space and are delegated to [8].

Similarly, in the presence of primary interference, the PDF and CDF of $S_{m,i}^{I,n}$ can be expressed as

$$f_{S_{m,i}^{I,n}}(x) = \frac{x\eta P_{n,i} + P_{m,i}(\eta + P_{n,i})}{(xP_{n,i} + P_{m,i})^2} \left(e^{\frac{\Psi_i}{P_{m,i}}} - 1\right) e^{-\frac{x\eta + \Psi_i}{P_{m,i}}} + \frac{\Psi_i}{x^3 P_{n,i}^2} \times e^{\frac{x\eta + \Psi_i}{xP_{n,i}}} \left[(\Psi_i + xP_{n,i})\Gamma\left(0, \left(\eta + \frac{\Psi_i}{x}\right)\left(\frac{1}{P_{n,i}} + \frac{x}{P_{m,i}}\right)\right) + \frac{xP_{n,i}(x^2\eta P_{n,i} - \Psi_i P_{m,i})}{(x\eta + \Psi_i)(xP_{n,i} + P_{m,i})} e^{-\left(\eta + \frac{\Psi_i}{x}\right)\left(\frac{1}{P_{n,i}} + \frac{x}{P_{m,i}}\right)}\right].$$

$$F_{S_{m,i}^{I,n}}(x) = 1 - \frac{\left(1 - e^{-\frac{\Psi_i}{P_{m,i}}}\right) e^{-\frac{x\eta}{P_{m,i}}}}{1 + \frac{xP_{n,i}}{P_{m,i}}} - \frac{\Psi_i}{xP_{n,i}} e^{\frac{\Psi_i}{xP_{n,i}} + \frac{\eta}{P_{n,i}}} \times \Gamma\left(0, \left(\eta + \frac{\Psi_i}{x}\right)\left(\frac{1}{P_{n,i}} + \frac{x}{P_{m,i}}\right)\right),$$

Finally, the desired expressions for the PDFs of $C_{m,i}^{I,n}$ and $C_{m,i}^{NI}$ can be obtained through the transformation of appropriately defined RVs as follows:

$$f_{C_{m,i}^{I,n}}(x) = \left|\frac{dy}{dx}\right| f_{S_{m,i}^{I,n}}(y)\Big|_{y=e^x-1} = e^x f_{S_{m,i}^{I,n}}(e^x - 1),$$
$$f_{C_{m,i}^{NI}}(x) = e^x f_{S_{m,i}^{NI}}(e^x - 1).$$

Recalling the SU capacity expression in (2), in the presence of N interfering PUs, there are two types of well known methods available to evaluate the distribution for sum of variates, namely, the characteristic function (CF) and the moment generating function (MGF) method. Unfortunately, employing these methods leads to intractable results and no explicit closed form expressions for the PDF and CDF of SU capacity in (2) can be achieved.

We will resort in this regard to an alternative approach. To sum up the rates for the cases of interference and no-interference, we will approximate the PDFs of $C_{m,i}^{I,n}$ and $C_{m,i}^{NI}$ using the Gamma distribution. There are several desirable properties of the Gamma distribution that are fit for approximating the PDFs of the variables $C_{m,i}^{I,n}$ and $C_{m,i}^{NI}$. First, the sum of Gamma distributed RVs with the same scale parameters is another Gamma distributed RVs. Second, the skewness and tail of distribution are similar for the whole range of interest and are determined by mean and variance [9]. Last but not least, Gamma distribution is a Type-III Pearson distribution, which is widely used in fitting *positive* RVs [9], [10]. In addition, since Gamma distribution is uniquely determined by its mean and variance, we will make use of the moment matching method to match the first two moments of the RV, namely its mean and variance.

Definition 1: Random variable X follows a Gamma distribution, $X \sim \mathcal{G}(\alpha, \beta)$ with scale and shape parameters $\beta > 0$ and $\alpha > 0$, respectively, if: $f_X(x) = \frac{x^{\alpha-1}\exp\left(-\frac{x}{\beta}\right)}{\beta^\alpha \Gamma(\alpha)} U(x)$, where $U(\cdot)$ represents the unit step function, and the Gamma function is defined as $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$.

Since the mean and variance of Gamma distribution are $\alpha\beta$ and $\alpha\beta^2$, respectively, matching the first two moments with the PDFs of $C_{m,i}^{I,n}$ and $C_{m,i}^{NI}$ leads to

$$\alpha_n^I = \frac{\left(\mathbb{E}\left[C_{m,i}^{I,n}\right]\right)^2}{\mathbf{var}\left[C_{m,i}^{I,n}\right]}, \quad \beta_n^I = \frac{\mathbf{var}\left[C_{m,i}^{I,n}\right]}{\mathbb{E}\left[C_{m,i}^{I,n}\right]},$$
$$\alpha^{NI} = \frac{\left(\mathbb{E}\left[C_{m,i}^{NI}\right]\right)^2}{\mathbf{var}\left[C_{m,i}^{NI}\right]}, \quad \beta^{NI} = \frac{\mathbf{var}\left[C_{m,i}^{NI}\right]}{\mathbb{E}\left[C_{m,i}^{NI}\right]},$$

for $n = 1, 2, \ldots, N$, and $\mathbf{var}(x)$ denotes the variance of x.

From [11], by employing the derived PDFs and CDFs of $S_{m,i}^{I,n}$ and $S_{m,i}^{NI}$, the average capacity of $C_{m,i}^{NI}$ and $C_{m,i}^{I,n}$ are expressed, respectively, as

$$\mathbb{E}\left[C_{m,i}^{NI}\right] = \int_0^\infty x f_{C_{m,i}^{NI}}(x)\mathrm{d}x = \int_0^\infty \log(1+x) f_{S_{m,i}^{NI}}(x)\mathrm{d}x = \Gamma\left(0, \frac{\eta}{P_{m,i}}\right) e^{\frac{\eta}{P_{m,i}}} \left(1 + \frac{e^{-\frac{\Psi_i}{P_{m,i}}}\eta}{\Psi_i - \eta}\right) + \frac{\Psi_i}{\eta - \Psi_i}\Gamma\left(0, \frac{\Psi_i}{P_{m,i}}\right).$$

and

$$\mathbb{E}\left[C_{m,i}^{I,n}\right] = \int_0^\infty \log(1+x) f_{S_{m,i}^{I,n}}(x)\mathrm{d}x = \frac{1 - e^{-\frac{\Psi_i}{P_{m,i}}}}{1 - \frac{P_{n,i}}{P_{m,i}}} \times \left(\Gamma\left(0, \frac{\eta}{P_{m,i}}\right) e^{\frac{\eta}{P_{m,i}}} - \Gamma\left(0, \frac{\eta}{P_{n,i}}\right) e^{\frac{\eta}{P_{n,i}}}\right) + \frac{\Psi_i}{P_{n,i}} e^{\frac{\eta}{P_{n,i}}} \int_0^\infty \Gamma\left(0, \left(\eta + \frac{\Psi_i}{x}\right) \times \left(\frac{1}{P_{n,i}} + \frac{x}{P_{m,i}}\right)\right) \frac{e^{\frac{\Psi_i}{xP_{n,i}}}}{x(1+x)}\mathrm{d}x,$$

The variance of $C_{m,i}^{I,n}$ is given by

$$\mathbf{var}\left[C_{m,i}^{I,n}\right] = \mathbb{E}\left[\left(C_{m,i}^{I,n}\right)^2\right] - \left(\mathbb{E}\left[C_{m,i}^{I,n}\right]\right)^2,$$

 ISBN: 978-1-61208-279-0

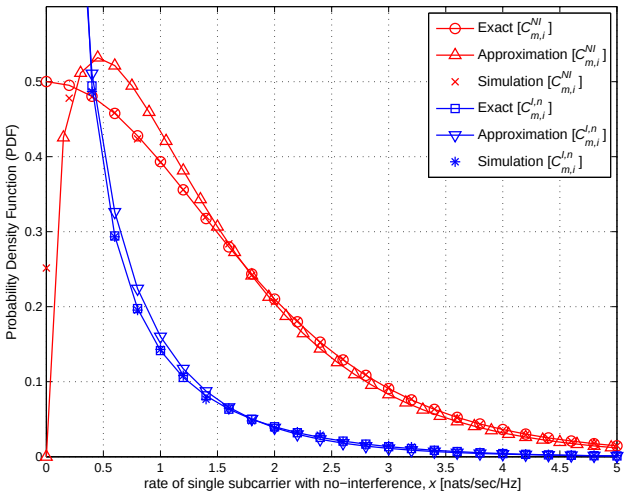


Figure 1. Comparison between the exact, approximation and simulation of $f_{C_{m,i}^{I,n}}(x)$ and $f_{C_{m,i}^{NI}}(x)$ using the PDF of Gamma distribution for $P_{m,i} = 20$ dB, $P_{n,i} = 10$ dB, $\Psi_i = 0$ dB and $\eta = 0.5$.

where the second moment of $C_{m,i}^{I,n}$ is expressed as

$$\mathbb{E}\left[\left(C_{m,i}^{I,n}\right)^2\right] = \int_0^\infty \left[\log(1+x)\right]^2 f_{S_{m,i}^{I,n}}(x)\mathrm{d}x$$
$$= \int_0^\infty \frac{2\log(1+x)}{1+x}\left[1 - F_{S_{m,i}^{I,n}}(x)\right]\mathrm{d}x$$
$$\simeq \sum_{j=1}^{N_p} w_j \frac{2\log(1+s_j)}{1+s_j}\left[1 - F_{S_{m,i}^{I,n}}(s_j)\right],$$

where the second equality is obtained by using integration by parts [11]. The resulting integral is estimated via Gauss-Chebyshev quadrature (GCQ), where the weights (w_j) and abscissas (s_j) are given by [12, Eqs. (22) and (23)], respectively. Similarly, the variance of $C_{m,i}^{NI}$ is expressed by adopting the same approach.

Based on the adopted Gamma approximation, the capacities are approximated as $C_{m,i}^{I,n} \sim \mathcal{G}\left(\alpha_n^I, \beta_n^I\right)$ and $C_{m,i}^{NI} \sim \mathcal{G}\left(\alpha^{NI}, \beta^{NI}\right)$.

In Figure 1, the exact and approximative expressions of $f_{C_{m,i}^{I,n}}(x)$ and $f_{C_{m,i}^{NI}}(x)$, including the simulations results, for different system parameters are depicted. It can be observed that the proposed approximation is very close to the exact results. Since both $C_{m,i}^{I,n}$ and $C_{m,i}^{NI}$ are i.i.d. for given k_{nm}, the conditional characteristic functions for the rate sums $\sum_{i=1}^{k_{nm}} C_{m,i}^{I}$ and $\sum_{i=1}^{k_{fm}} C_{m,i}^{NI}$ can be expressed as follows

$$\Phi_{C_m^{I,n}}(\omega|k_{nm}) = \left(\Phi_{C_{m,i}^{I,n}}(\omega)\right)^{k_{nm}} = \left(1 - j\omega\beta_n^I\right)^{-\alpha_n^I k_{nm}},$$
$$\Phi_{C_m^{NI}}(\omega|k_{nm}) = \left(\Phi_{C_{m,i}^{NI}}(\omega)\right)^{k_{fm}} = \left(1 - j\omega\beta^{NI}\right)^{-\alpha^{NI} k_{fm}},$$

where $\Phi_{C_{m,i}^{I,n}}(\omega|k_{nm})$ and $\Phi_{C_{m,i}^{NI}}(\omega|k_{nm})$ are the characteristic functions of $f_{C_{m,i}^{I,n}}(x|k_{nm})$ and $f_{C_{m,i}^{NI}}(x|k_{nm})$, respectively. Using the property of the Gamma distribution that the sum of i.i.d. Gamma distributed RVs, with the same scale parameters (β) is another Gamma distributed RV, the conditional PDFs take the form:

$$f_{C_m^{I,n}|k_{nm}}(x|k_{nm}) = \mathcal{G}\left(\alpha_n^I k_{nm}, \beta_n^I\right),$$
$$f_{C_m^{NI}|k_{nm}}(x|k_{nm}) = \mathcal{G}\left(\alpha^{NI} k_{fm}, \beta^{NI}\right). \quad (4)$$

In (2), even though the conditional PDFs of $C_m^{I,n}$ and C_m^{NI} are calculated, to find the PDF expression for C_m, one first needs to evaluate the PDF of C_m^I, and then the PDF of its sum with C_m^{NI}. Notice that there are $N+1$ terms in (2), and each follows a Gamma distribution where the shape (α) and scale (β) parameters can be arbitrary. Therefore, the aforementioned property of Gamma distribution for a sum of Gamma variates cannot be employed anymore.

The expression for the PDF of a sum of Gamma RVs was obtained by Moschopoulos in [13], where a mathematically tractable solution that does not restrict the scale and shape parameters to be integer-valued or all distinct is presented. Therefore, the following theorem will be used next.

Theorem 1 (Moschopoulos, 1985): Let $\{X_s\}_{s=1}^{\mathcal{S}}$ be independent but not necessarily identically distributed Gamma variates with parameters α_s and β_s, respectively, then the PDF of $Y = \sum_{s=1}^{\mathcal{S}} X_s$ can be expressed as

$$f_Y(y) = \prod_{s=1}^{\mathcal{S}} \left(\frac{\beta_1}{\beta_s}\right)^{\alpha_s} \sum_{k=0}^{\infty} \frac{\delta_k y^{\sum_{s=1}^{\mathcal{S}} \alpha_s + k - 1} \exp\left(-\frac{y}{\beta_1}\right)}{\beta_1^{\sum_{s=1}^{\mathcal{S}} \alpha_s + k} \Gamma\left(\sum_{s=1}^{\mathcal{S}} \alpha_s + k\right)} U(y), \quad (5)$$

where $\beta_1 = \min_s\{\beta_s\}$, and the coefficients δ_k can be obtained recursively by the formula

$$\delta_k = \frac{1}{k+1} \sum_{i=1}^{k+1} \left[\sum_{j=1}^{\mathcal{S}} \alpha_j \left(1 - \frac{\beta_1}{\beta_j}\right)^i\right] \delta_{k+1-i}$$

where $\delta_0 = 1$, and for $k = 0, 1, 2, \ldots$.

Proof: See [13]. ■

The Moschopoulos PDF provides a tractable representation for the sum of Gamma variates in terms of a single Gamma series via a recursive formula to evaluate iteratively the representation coefficients. This approach is applicable for any arbitrary shape parameters $\{\alpha_s\}_{s=1}^{\mathcal{S}}$ and scale parameters $\{\beta_s\}_{s=1}^{\mathcal{S}}$ including the possibility of having some of the parameters identical. Notice that in the considered communication system, with some probability the transmit power of PUs $P_{n,i}$ for $n = 1, \ldots, N$, can be the same, which means that the corresponding α_n^I and β_n^I are the same. Such a set-up might arise when the PUs are at the same distance from their corresponding common PBS.

 ISBN: 978-1-61208-279-0

The CDF of Y can be obtained from the PDF as $F_Y(y) = \int_{-\infty}^{y} f_Y(x)\mathrm{d}x$. Therefore,

$$F_Y(y) = \prod_{s=1}^{\mathcal{S}} \left(\frac{\beta_1}{\beta_s}\right)^{\alpha_s} \sum_{k=0}^{\infty} \frac{\delta_k}{\beta_1^{\sum_{s=1}^{\mathcal{S}} \alpha_s + k} \Gamma\left(\sum_{s=1}^{\mathcal{S}} \alpha_s + k\right)} \times \int_0^y x^{\sum_{s=1}^{\mathcal{S}} \alpha_s + k - 1} \exp\left(-\frac{x}{\beta_1}\right) \mathrm{d}x. \tag{6}$$

Notice that the interchange of summation and integration operators is justified due to the uniform convergence of (5). From [14], we can simplify (6) by using $\int_0^u x^{\nu-1} e^{-\mu x}\mathrm{d}x = \mu^{-\nu}\gamma(\nu, \mu u)$ for $\Re[\nu > 0]$ [15, pg. 346, Sec. 3.381, Eq. 1], where $\gamma(\cdot,\cdot)$ is the lower incomplete Gamma function and is defined as $\gamma(x, y) = \int_0^y t^{x-1} e^{-t} dt$. Hence,

$$\begin{aligned} F_Y(y) &= \prod_{s=1}^{\mathcal{S}} \left(\frac{\beta_1}{\beta_s}\right)^{\alpha_s} \sum_{k=0}^{\infty} \delta_k \frac{\gamma\left(\sum_{s=1}^{\mathcal{S}} \alpha_s + k, \frac{y}{\beta_1}\right)}{\Gamma\left(\sum_{s=1}^{\mathcal{S}} \alpha_s + k\right)} \\ &= \prod_{s=1}^{\mathcal{S}} \left(\frac{\beta_1}{\beta_s}\right)^{\alpha_s} \sum_{k=0}^{\infty} \delta_k \mathcal{P}\left(\sum_{s=1}^{\mathcal{S}} \alpha_s + k, \frac{y}{\beta_1}\right), \end{aligned} \tag{7}$$

where $\mathcal{P}(\cdot,\cdot)$ is referred to as the regularized (also termed normalized) incomplete Gamma function: $\mathcal{P}(a, z) = \frac{\gamma(a,z)}{\Gamma(a)} = 1 - \frac{\Gamma(a,z)}{\Gamma(a)}$. Based on the required accuracy, one may consider the first h, i.e., $k = h - 1$, terms in the sum series (5). The expression for truncation error is given in [13].

Recall that from (2) and (4), we have to determine the PDF of the sum $C_m^{I,1} + C_m^{I,2} + \cdots + C_m^{I,N} + C_m^{NI}$, for a given number of subcarrier collisions $\mathbf{k}_m = [k_{1m}, k_{2m}, \ldots, k_{Nm}, k_{fm}]$. Recall also that C_m^I and C_m^{NI} are Gamma distributed and independent but not necessarily identical. Therefore, the conditional PDF of their sum can be obtained by means of *Theorem* 1 as given in (8). Equation (8) displayed at the top of the next page describes the sought result. where $\beta_{\min} = \min\{\beta_1^I, \beta_2^I, \ldots, \beta_N^I, \beta^{NI}\}$, and the coefficients δ_k are obtained recursively:

$$\delta_k = \frac{1}{k+1} \sum_{i=1}^{k+1} \left[\sum_{j=1}^{N} \alpha_i^I k_{jm} \left(1 - \frac{\beta_{\min}}{\beta_j^I}\right)^i + \alpha^{NI} k_{fm} \left(1 - \frac{\beta_{\min}}{\beta^{NI}}\right)^i \right] \delta_{k+1-i} \quad \text{for } k = 0, 1, 2, \ldots$$

where $\delta_0 = 1$.

Now, the PDF of C_m can be determined by averaging over the PMF of subcarrier collisions:

$$\begin{aligned} f_{C_m}(x) &= \sum_{\mathbf{k}_m} f_{C_m, \mathbf{K}_m}(x, \mathbf{k}_m) \\ &= \sum_{\mathbf{k}_m} f_{C_m|\mathbf{K}_m}(x|\mathbf{k}_m) p(\mathbf{k}_m). \end{aligned} \tag{10}$$

Plugging (1) and (8) into (10) yields the sought PDF in (9).

The outage probability is a often used performance metric in channels subject to fading conditions. Hence, herein paper we will determine the outage probability of SU capacity in terms of $P_{C_m}^{\text{out}}(\varphi_{\text{th}}) = Pr\left(C_m < \varphi_{\text{th}}\right) = \int_0^{\varphi_{\text{th}}} f_{C_m}(x)\mathrm{d}x$, which represents the CDF of the SU capacity over the outage threshold φ_{th} [dB].

Using (7) and (9), the CDF of C_m takes the form:

$$\begin{aligned} F_{C_m}(x) = \sum_{k_{1m}} \sum_{k_{2m}} \cdots \sum_{k_{Nm}} \sum_{k_{fm}} \Bigg\{ &\binom{F_f}{k_{fm}} \binom{F}{F_m^S}^{-1} \\ &\times \prod_{n=1}^{N} \binom{F_n^P}{k_{nm}} \left(\frac{\beta_{\min}}{\beta^{NI}}\right)^{\alpha^{NI} k_{fm}} \prod_{n=1}^{N} \left(\frac{\beta_{\min}}{\beta_n^I}\right)^{\alpha_n^I k_{nm}} \\ &\times \sum_{k=0}^{\infty} \delta_k \mathcal{P}\left(\sum_{n=1}^{N} \alpha_n^I k_{nm} + \alpha^{NI} k_{fm} + k, \frac{x}{\beta_{\min}}\right) \Bigg\}. \end{aligned}$$

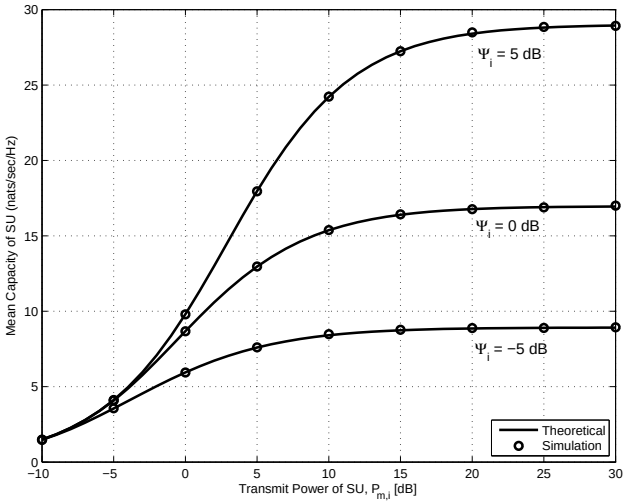


Figure 2. SU mean capacity versus the transmit power $P_{m,i}$ with different IT Ψ_i values for $F_m^S = 20$, $F_n^P = 30$, $F = 128$ and $P_{n,i} = 10$ dB.

IV. Computer Simulations

The influence of the secondary user peak transmit power $P_{m,i}$ (in dB) on its average capacity (in nats per second per hertz) is illustrated for different values of IT Ψ_i in Fig. 2. It turns out that the cognitive user's average capacity gets saturated after a certain value of peak SU transmit power because of the IT constraint. Fig. 2 corroborates the fact that the analytical results agree well with the simulation results.

 ISBN: 978-1-61208-279-0

$$f_{C_m|\mathbf{K}_m}(x|\mathbf{k}_m) = \left(\frac{\beta_{\min}}{\beta^{NI}}\right)^{\alpha^{NI}k_{fm}} \prod_{n=1}^{N}\left(\frac{\beta_{\min}}{\beta_n^I}\right)^{\alpha_n^I k_{nm}} \sum_{k=0}^{\infty} \frac{\delta_k x^{\sum_{n=1}^{N}\alpha_n^I k_{nm}+\alpha^{NI}k_{fm}+k-1}\exp\left(-\frac{x}{\beta_{\min}}\right)U(x)}{\beta_{\min}^{\sum_{n=1}^{N}\alpha_n^I k_{nm}+\alpha^{NI}k_{fm}+k}\,\Gamma\left(\sum_{n=1}^{N}\alpha_n^I k_{nm}+\alpha^{NI}k_{fm}+k\right)}, \tag{8}$$

$$\begin{aligned} f_{C_m}(x) = &\sum_{k_{1m}}\sum_{k_{2m}}\cdots\sum_{k_{Nm}}\sum_{k_{fm}}\left\{\binom{F_f}{k_{fm}}\binom{F}{F_m^S}^{-1}\prod_{n=1}^{N}\binom{F_n^P}{k_{nm}}\left(\frac{\beta_{\min}}{\beta^{NI}}\right)^{\alpha^{NI}k_{fm}}\prod_{n=1}^{N}\left(\frac{\beta_{\min}}{\beta_n^I}\right)^{\alpha_n^I k_{nm}}\right. \\ &\left.\times\sum_{k=0}^{\infty}\frac{\delta_k x^{\sum_{n=1}^{N}\alpha_n^I k_{nm}+\alpha^{NI}k_{fm}+k-1}\exp\left(-\frac{x}{\beta_{\min}}\right)}{\beta_{\min}^{\sum_{n=1}^{N}\alpha_n^I k_{nm}+\alpha^{NI}k_{fm}+k}\,\Gamma\left(\sum_{n=1}^{N}\alpha_n^I k_{nm}+\alpha^{NI}k_{fm}+k\right)}U(x)\right\}. \end{aligned} \tag{9}$$

The plots in Figure 2 are in the presence of a single PU, i.e., $n \in [1, N]$, and unit variance AWGN ($\eta = 1$). The number of subcarriers in sets F, F_m^S and F_n^S is selected arbitrarily. Fig. 2 highlights also the fact that the saturation level of capacity increases as the IT constraint relaxes, and the capacity keeps growing until a saturation level as the transmit power of SU increases. However, the capacity gains due to the relaxation in the IT constraint disappears in the low SU transmit power regime.

V. Conclusions

This paper assessed the capacity of a secondary (cognitive) user in a random access OFDM-based cognitive radio system with spectrum sharing features such as random subcarrier allocation and absence of spectrum sensing information. The adopted model for the number of subcarrier collisions in the presence of multiple interfering primary users is the general multivariate hypergeometric distribution. The PDF and CDF expressions of the secondary user capacity over a Rayleigh fading channel are derived. It turns out that the closed-form expression for the instantaneous secondary user capacity over Rayleigh channel fading is intractable. Therefore, a Gamma approximation of the secondary user capacity is obtained by employing the moment matching method and the concept of Moschopoulos PDF representation. The work conducted in this paper subscribes along the lines of our preliminary results [8], and we are hoping to extend these results to a more general OFDM-based cognitive radio network that assumes an arbitrary number of primary and secondary users, and general channel fading conditions.

Acknowledgment

This work was made possible by the support offered by QNRF-NPRP grants 09-341-2128, 4-1293-2-513 and 5-250-2-087.

References

[1] S. Haykin, “Cognitive radio: Brain-empowered wireless communications,” IEEE Journal on Selected Areas in Communications, vol. 23, no. 2, pp. 201–220, February 2005.

[2] T. W. Ban, W. Choi, B. C. Jung, and D. K. Sung, “Multi-user diversity in a spectrum sharing system,” IEEE Transactions on Wireless Communications, vol. 8, no. 1, pp. 102–106, January 2009.

[3] D. Dongliang, Y. Liuqing, and J. C. Principe, “Cooperative diversity of spectrum sensing for cognitive radio systems,” IEEE Transactions on Signal Processing, vol. 58, no. 6, pp. 3218–3227, June 2010.

[4] Z. Tian and G. B. Giannakis, “Compressed sensing for wideband cognitive radios,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), IEEE Press, 2007, pp. IV-1357–IV-1360.

[5] T. Yucek and H. Arslan, “A survey of spectrum sensing algorithms for cognitive radio applications,” IEEE Communications Surveys and Tutorials, vol. 11, no. 1, pp. 116–130, First Quarter 2009.

[6] R. Zhang, “On peak versus average interference power constraints for protecting primary users in cognitive radio networks,” IEEE Transactions on Wireless Communications, vol. 8, no. 4, pp. 2112–2120, April 2009.

[7] R. Zhang and Y.-C. Liang, “Investigation on multiuser diversity in spectrum sharing based cognitive radio networks,” IEEE Communications Letters, vol. 14, no. 2, pp. 133–135, February 2010.

[8] S. Ekin, M. M. Abdallah, K. A. Qaraqe, and E. Serpedin, “Random Subcarrier Allocation in OFDM-Based Cognitive Radio Networks,” IEEE Transactions on Signal Processing, vol. 60, no. 9, pp. 4758–4774 , September 2012.

[9] J. Wagner, Y.-C. Liang, and R. Zhang, “On the balance of multiuser diversity and spatial multiplexing gain in random beamforming,” IEEE Transactions on Wireless Communications, vol. 7, no. 7, pp. 2512–2525, July 2008.

 ISBN: 978-1-61208-279-0

[10] S. Al-Ahmadi and H. Yanikomeroglu, "On the approximation of the generalized-k distribution by a gamma distribution for modeling composite fading channels," IEEE Transactions on Wireless Communications, vol. 9, no. 2, pp. 706–713, February 2010.

[11] H. A. Suraweera, P. J. Smith, and M. Shafi, "Capacity limits and performance analysis of cognitive radio with imperfect channel knowledge," IEEE Transactions on Vehicular Technology, vol. 59, no. 4, pp. 1811–1822, May 2010.

[12] F. Yilmaz and M.-S. Alouini, "An MGF-based capacity analysis of equal gain combining over fading channels," Proc. IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC 2012), IEEE Press, 2010, pp. 945–950.

[13] P. G. Moschopoulos, "The distribution of the sum of independent gamma random variables," Annals of Institute of Statistical Mathematics (Part A), vol. 37, no. 1, pp. 541–544, 1985.

[14] H. Suraweera, P. Smith, and J. Armstrong, "Outage probability of cooperative relay networks in Nakagami-m fading channels," IEEE Communications Letters, vol. 10, no. 12, pp. 834–836, Dec. 2006.

[15] I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series, and Products, 7th ed., A. Jeffrey and D. Zwillinger, Eds., Academic Press, 2007.

 ISBN: 978-1-61208-279-0

Spectral Occupancy Measurements in Rural and Urban Environments: Analysis and Comparison

Alexandru Marţian, Călin Vlădeanu, Octavian Fratu, Ion Marghescu
Telecommunications Department
Politehnica University of Bucharest
Bucharest, Romania
E-mail: martian@radio.pub.ro, calin@comm.pub.ro, ofratu@elcom.pub.ro, marion@comm.pub.ro

Safwan El Assad
IETR Laboratory, UMR CNRS 6164, Image team
École d'Ingénieurs de l'Université de Nantes
Nantes, France
E-mail: safwan.elassad@univ-nantes.fr

Abstract— **In order to enable the coexistence of an ever-increasing number of communication systems in a limited amount of frequency, the traditional static frequency allocation is not the best solution any longer. Cognitive Radio (CR) technology, based on a dynamic spectrum access, is a possible solution for improving the efficiency of spectrum usage. A first step in order to identify the frequency bands that are more suitable for opportunistic usage is to evaluate the degree in which licensed bands are currently used. Although some measurement campaigns have already been carried out, most of them were done in urban environments and only a few in other locations. This paper presents results of two measurement campaigns conducted in Romania both in urban and rural environments, covering the frequency range from 25 MHz up to 3.4 GHz. The results are confronted with the frequency allocation table published by the national authority for radio communications and a comparison and analysis of the obtained data are being made.**

Keywords - spectral occupancy; cognitive radio; dynamic spectrum access; energy detection; measurement campaign.

I. INTRODUCTION

During the last decades, the demand for more radio spectrum increased with the development of wireless communications. With the deployment of more wireless communications systems, most of the available spectrum has been statically allocated. Therefore, many countries are facing the problem of spectrum insufficiency. Nevertheless, measurement studies have revealed that most of the allocated spectrum experiences low utilization efficiency. These two facts motivate the introduction of dynamic spectrum access, which allows secondary users to reuse/share the same radio spectrum originally allocated to the primary (licensed) users.

Cognitive radio (CR) technology is the key technology that enables a system to use spectrum in a dynamic manner. The CR term was coined by J. Mitola III in [1], defining a wireless communication system that allows spectrum sharing over a wide frequency range and that is able to handle multiple radio access technologies. In order to avoid any interference to the primary system, the CR equipment should include the following functionalities: frequency-agility and re-configuration of radios, spectrum sensing, and spectrum management.

Prior to developing standards for CR applications, there is a need for a thorough investigation of potential spectral regions to be used. Therefore, it is important to gather spectral investigation results from as many as possible different geographical regions and scenarios. These results will facilitate the CR standards development, to make the CR devices work under several circumstances, with different spectral regulations.

Several measurement campaigns concerning spectrum occupancy were conducted worldwide [2]-[10], most of them were carried out the in the USA [2]-[3] and only a few in other locations worldwide, including Singapore [5], Germany [6], New Zeeland [7], Spain [8] and Italy [9], in urban or suburban scenarios. Results of a measurement campaign conducted in Chicago, USA showed a mean occupancy as low as 17.4% in the frequency band 30 to 3000 MHz [2]. Studies were also targeted at narrower frequency bands, like the public safety ones, and the benefits of cooperative sensing were highlighted [3]. The difference between indoor and outdoor locations was discussed in [5] based on measurements performed in Aachen, Germany. The study in [9], based on measurements performed in Spain and Italy, focuses on determining the hidden noise margin in order to find out if cognitive devices are able to detect and distinguish between empty and occupied TV channels. Recently, after a thorough investigation of real radio communications systems, accurate statistical models for the time-domain spectral occupancy were proposed in [10].

The paper is organized as follows. Section II contains a description of the equipments used to perform our measurements. The methodology and the obtained results of the measurements are presented in Section III. In Section IV, the measurement results are analyzed from a CR perspective. Finally, in Section V, the conclusions are drawn and future work aspects are presented.

II. MEASUREMENT LOCATION, SETUP AND EQUIPMENT

The measurements for an urban environment were performed in Bucharest from the top of the main building of

 ISBN: 978-1-61208-279-0

our Department (GPS location: latitude 44°26'01" N, longitude 26°03'27" E, MSL altitude 150 m, relative altitude 30 m). The location is excellent for such a measurement campaign, having direct line of sight with several FM transmitters, Analog and DVB-T TV transmitters, GSM and UMTS base stations and several other stations The headquarters of the governmental agency for special telecommunications is also located just a few hundreds of meters away from the measurement location.

The small village of Maneciu (GPS location: latitude 45°18'49" N, longitude 25°59'38" E, MSL altitude 584 m, relative altitude 10 m) was chosen for the measurements in a rural environment. The village is located in a hilly region approximately 100 km away from Bucharest and at least 40 km away from any other big city. A map of the southeastern part of Romania, having highlighted the measurement locations is given in Figure 1.

The antenna used for collecting the measurement data is a wideband discone antenna (Sirio SD3000N), mounted on the building terrace. The antenna has an omnidirectional pattern in the horizontal plane and was connected using a low-loss RF cable to a high performance signal analyzer (Anritsu MS2690A - 50 Hz to 6 GHz). Images showing the equipment that was used during the measurement campaigns can be found in [11].

The measurements covered the frequency range from 25 MHz to 3400 MHz, the whole band being divided into 14 sub-bands according to the type of service and the bandwidth of the allocated signal.

The measurement data was collected from the spectrum analyzer using a remote mode, the necessary commands were generated using a notebook connected through the network interface. Data was acquired for each of the 14 spectrum sub-bands at intervals of around 70 seconds, resulting more than 1000 samples for each of the sub-bands for each day.

A list containing values of the parameters used to configure the equipment in order to perform the measurements is given in Table I.

TABLE I. SPECTRUM ANALYZER CONFIGURATION

Parameter	Value	
Frequency bands	1.25-230 MHz	8.1525-1710 MHz
	2.230-400 MHz	9.1710-1880 MHz
	3.400-470 MHz	10.1880-2200 MHz
	4.470-766 MHz	11.2200-2400 MHz
	5.766-880 MHz	12.2400-2500 MHz
	6.880-960 MHz	13.2500-2690 MHz
	7.960-1525 MHz	14.2690-3400 MHz
Resolution/video bandwidth (RBW/VBW)	300 kHz / 300 kHz (bands 2, 3, 5, 6, 8, 9, 11, 12, 13) 1 MHz / 1MHz (bands 1, 4, 7, 10, 14)	
Sweep time	5 ms	
Reference level	0 dBm	
Attenuation	10 dB	
Detection type	Pos & Neg	
Trace points	10001	

The MathWorks MATLAB software environment was used for collecting, processing and analyzing the data obtained during the measurement campaign.

III. MEASUREMENT RESULTS

Several sensing methods can be used in order to decide if a certain frequency band is available for opportunistic access: energy detection, matched filtering, cyclostationary feature detection, eigenvalue detection, wavelet detection [12]-[15]. In order to evaluate the spectral occupancy in the several frequency bands presented in Table I using the data collected during the measurement campaigns we chose the energy detection method.

The energy detection method provides an optimal detection in cases where the primary user signal is unknown, as it does not require any a priori knowledge about it. Some other advantages offered by this method are low computational costs and a short necessary sensing time. The main drawbacks of this method are that the results obtained are highly susceptible to changes in the background noise and interference level and that it cannot distinguish primary users from secondary ones. However, research has been done, like the one described in [12] in order to improve the detection performance of such a method by dynamically adjusting the threshold value.

One of the main challenges when using the method is a proper selection of the energy threshold. If the value used for the threshold is too high, weak signals will be treated as noise and this would result in an underestimation of the actual occupancy. On the other hand, choosing a too low value for the threshold will increase the false alarm probability caused by high-power noise samples and this would cause an overestimation of the actual occupancy.

In order to estimate the noise level, several methods are available. Although the most simple one would imply a measurement using a matched load (of 50 Ω in case of our configuration) instead of the measurement antenna, this would imply that the evaluation of the noise level would be made at a different moment of time from the one when the measurements are being made. In this case, the accuracy of the measurement would not be acceptable, and moreover the values obtained could not be dynamically adjusted to reflect possible changes in the radio environment. To overcome these disadvantages we determined the noise level using a sliding window in the frequency domain, having an adaptive width. An algorithm was used to determine the most wide frequency interval without any active signal for each of the 14 frequency bands described in Table I, and a mean of the values for this interval would indicate the current noise level for the corresponding band.

To mitigate the effects of high-power noise samples in order to decrease the false alarm probability, a second sliding window with a calculated width of 100 kHz was used in order to mean out such samples.

In Table II, an example calculation was made for the 470 to 766 MHz frequency band by choosing several different values for the false alarm probability.

 ISBN: 978-1-61208-279-0

TABLE II. INFLUENCE OF THRESHOLD VALUE OVER MEASURED OCCUPANCY FOR THE 470-766 MHZ FREQUENCY BAND (URBAN ENVIRONMENT, HOUR INTERVAL 14-22, WORKDAY)

False alarm probability (%)	Spectral occupancy (%)	Threshold level (dBm)	Threshold above noise level (dB)
1	34,21	-84.82	4,80
3	38,52	-85.67	3,95
5	40,43	-85.95	3,67
7	42,03	-86.13	3,49
10	**44,24**	**-86.33**	**3,29**
15	47,56	-86.55	3,07
20	48,92	-86.64	2,98

The threshold value was chosen in order to respect a chosen false alarm probability of 10% for all the frequency bands that were considered during the measurement campaigns. As expected, when the value chosen for the false alarm probability is increased, the value of the necessary threshold will decrease. As a result, the measured spectral occupancy will have higher values, as signals having levels close to the noise level will be taken into account.

As it can be noticed, having a 19 % variation in the false alarm probability implies also almost a 15% variation in the determined occupancy. The value of 10% for the false alarm probability was highlighted, as it is the value that was used during the measurement campaigns.

Table III contains the measured spectrum occupancies for both the urban and rural environments for the 25 MHz to 3.4 GHz frequency band. The listed results were obtained by using a value of 10% for the false alarm probability for the threshold for each of the 14 different frequency sub-bands.

Occupancy results were calculated for both the urban and the rural areas using data collected over intervals of 48 hours, one working day and one week-end day.

IV. MEASUREMENT ANALYSIS

Figures 2 and 3 contain spectrum occupancy results for some of the 14 frequency sub-bands analyzed during the urban environment measurement campaign. In case of sub-band 470-766 MHz (Figure 2) different metrics are used in order to express the degree of spectral occupancy. A first sub-figure contains the maximum, minimum and mean power spectral density value for an 8-hour observation interval. In a second sub-figure the occupancy duty cycle is represented, expressing the amount of time (in %) in with each of the frequencies from the respective frequency band was occupied or not. A third sub-figure is a histogram showing a map of the instantaneous spectral occupancy for the whole 8-hour interval. In case of Figure 4 the maximum, minimum and mean power spectral density value are given for four different sub-bands.

The frequency spectrum below 1 GHz shows the highest occupancy values from the whole frequency domain that was covered during our measurement campaigns, both in the urban and rural environments. Nevertheless, in both cases the calculated values are below 50% (37.23% in the urban area, 19.19% in the rural area), which means that even in this frequency range there are bands with some potential for cognitive radio applications. Another aspect worth being noticed is that for this frequency interval the difference between the two environments is the most significant one, the occupancy being almost double in case of the urban area. The reason for this difference is the lower density of broadcast radio and TV stations and a lower number of frequency channels used in case of the GSM mobile networks in the rural area.

TABLE III. SPECTRUM OCCUPANCY IN BOTH URBAN AND RURAL ENVIRONMENTS FOR THE 25-3400 MHZ FREQUENCY RANGE

Frequency range (MHz)	**Possible applications according to TNABF [16]**	**Measured Occupancy (%)**		**Mean Occupancy (%)**			
		Urban environment	Rural environment	Urban environment	Rural environment	Urban environment	Rural environment
25 - 230	FM radio, Aero/Marine, Fixed/Mobile, Military, other applications	48.64	28.67	37.23	19.19	21.00	14.19
230 - 400	Military, Mobile	17.82	13.25				
400 - 470	Analogue/Digital Terrestrial Mobile, Meteorology, other applications	42.83	23.43				
470 - 766	Analogue TV, DVB-T	43.06	11.42				
766 - 880	Military, TV, DVB-T, Cordless, Military, other applications	17.63	11.10				
880 - 960	Mobile communication systems :GSM, E-GSM, Military	50.85	44.19				
960 - 1525	Aero/Naval, Navigation, Radar, Military, Radio astronomy	10.89	10.11	15.89	11.15		
1525 - 1710	Satellite Mobile, Military, Meteorology	10.91	10.83				
1710 - 1880	Mobile communication systems :GSM 1800, other applications	28.58	14.64				
1880 - 2200	Mobile communication systems :UMTS/IMT 2000, DECT, other applications	20.87	11.34				
2200 - 2400	SAP/SAB, Military	17.26	17.04	13.64	13.44		
2400 - 2500	ISM, RFID, RLAN, other applications	17.66	15.90				
2500 - 2690	Mobile communication systems :UMTS/IMT 2000, Military	21.13	21.01				
2690 - 3400	Military, Radar, Navigation, Meteorology, other applications	10.05	10.06				

 ISBN: 978-1-61208-279-0

The frequency band 25 to 230 MHz exhibits a quite high occupancy degree especially in the urban environment, mainly because of (Figure 3a). The occupancy results in case of this particular frequency band during the whole measurement period, divided in 6 time intervals, are given in Table IV.

The frequency band 470-766 MHz (Figure 2) is licensed for analog and digital TV broadcasting. Although due to European regulations analog TV broadcasting in Romania was initially intending to be completely replaced by digital one beginning with the 1st of January 2012, things are not quite there yet. In Figure 2 several analog TV broadcasting stations can be noticed in the urban area (the level for the station located on 506 MHz is higher than average, as the broadcast antenna is located on the same building from where the measurements were performed). This is also the frequency band where the biggest difference between the occupancy measured in the two environments was noticed (43.06% in the urban area compared to 11.42% in the rural one). The reason is mainly the DVB-T broadcasting that is for the moment performed only in the area surrounding Bucharest. The corresponding signals can be found in Figure 2 on the 546 MHz carrier frequency (channel 30) and on the 738 MHz carrier frequency (channel 54). Considering the fact that the spectral efficiency of the digital TV broadcasting is much higher than in case of the analog one, once the analog TV stations will cease broadcasting a drop in the spectral occupancy in this frequency band is to be expected. The Romanian National Authority for Management and Regulation in Communications (ANCOM) established 2013 as a new term for ceasing the analog TV broadcasting, so new measurements will have to be conducted in order to evaluate spectral occupancy in this frequency band afterwards. This frequency band is also targeted by several new standards based on dynamic spectrum access (including 802.22 and 802.11af).

The lowest spectral occupancy below 1 GHz was obtained in the 766 to 880 MHz band (Figure 3b), however most of this band is for the moment licensed for military applications. In this frequency band, only a test DVB-T broadcast station (carrier frequency 778 MHz, channel 59) was detected in the urban area during the measurement campaign, although other possible applications are allowed according to the National Table for Frequency Allocation in Romania [16].

The 880-960 MHz (Figure 3c) and 1710-1880 MHz are licensed for the GSM 900 and 1800 mobile communication system. In the frequency bands corresponding to the downlink direction (925-960 MHz for GSM 900 and 1805-1880 MHz for GSM 1800) a higher power level was measured, as the transmit power employed is considerably higher than the one used in case of mobile stations. In case of the rural environment, a significant difference in the degrees of occupancy was obtained in this frequency band depending on the time interval, as described in Table V.

In order to prove the reason for such a big difference in the degree of occupancy, in Figure 1 are pictured the histograms for the three 8-hour intervals listed in Table V.

As it can be clearly seen, a certain number of frequency channels used during the day (05:00 to 18:00) are turned off during nighttime, probably because of lack of traffic during off-work hours. Although in frequency bands corresponding to the uplink (880-915 MHz and 1710-1785 MHz) the measured occupancy was extremely low in both environments, it should be noted the measurement locations and the low transmit power of mobile stations might cause an underestimation of the real occupancy. CR equipment that are designed to function in these frequency bands should have a detection mechanism capable of recognizing low-power signals with energy close to the noise floor, in order to avoid interference to primary users active in the area.

The overall spectrum occupancy measured for the frequency bands above 1 GHz was extremely low for both environments (mean occupancies of less than 20% in both cases). As it can be noticed from Table III, the differences between the values obtained in the two areas are very small, excepting two bands: the 1710-1880 MHz band licensed for GSM 1800 and the 1880-2200 MHz band (Figure 3d) licensed for UMTS. In both cases because of the lower population density from the rural area the network operators use less frequency channels, this being the reason for the lower occupancy measured in the rural environment. Considering the very low spectrum occupancy degree, most of these frequency bands are potential candidates for CR applications.

TABLE IV. INFLUENCE OF DAY PERIOD OVER MEASURED OCCUPANCY FOR THE FREQUENCY BAND 25-230 MHZ (URBAN ENVIRONMENT)

Observation interval (hours : minutes)	Measured occupancy (%)
06:00 – 14:00 (workday)	47,68
14:00 – 22:00 (workday)	47,39
22:00 – 06:00 (workday)	47,27
06:00 – 14:00 (weekend)	51,06
14:00 – 22:00 (weekend)	50,22
22:00 – 06:00 (weekend)	48,25

TABLE V. INFLUENCE OF DAY PERIOD OVER MEASURED OCCUPANCY FOR THE FREQUENCY BAND 880-960 MHZ (RURAL ENVIRONMENT)

Time interval (hours : minutes)	Measured spectral occupancy (%)
06:00 – 14:00 (week-end day)	47,22
14:00 – 22:00 (week-end day)	43,98
22:00 – 06:00 (week-end day)	40,60

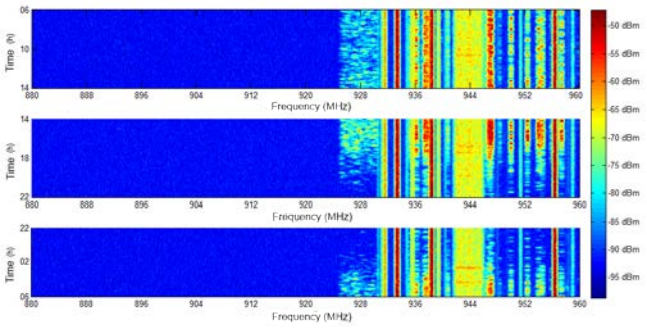


Figure 1. Histograms for the frequency band 880-960 MHz for three intervals of eight hours each (rural environment).

The ISM band 2400 to 2500 MHz is a very good opportunity for testing CR prototype devices, as the measured occupancy is quite low in both cases (17.66% in the urban environment and 15.90% in the rural area) and there are multitudes of commercially available hardware devices that are able to operate in this frequency range.

V. CONCLUSION AND FUTURE WORK

Results obtained during the measurement campaign conducted in both urban (Bucharest) and rural (Maneciu) environments in Romania clearly indicate that several frequency bands allow opportunistic access for future CR applications, especially in the rural environment and for the frequency bands above 1 GHz. The analyzed frequency range was 25 MHz to 3.4 GHz, and the mean occupancy ratio over the whole band was as low as 21.00% in the urban environment and 14.19% in the rural environment.

It is difficult to make a direct comparison between the results obtained during the different measurements campaigns mentioned in Section I and the measurement campaigns performed in Romania. The main reasons are the differences that can be found in the measurement setup (frequency intervals, spectrum analyzer bandwidth) and in the signal processing section (imposed false alarm probability). However, a common conclusion that can be drawn from the measurements is that the spectral occupancy degree that can be obtained by using a 'classical' static allocation approach is far from being optimal. The dynamic spectrum access proposed by technologies such as cognitive radio promises to bring substantial improvements in terms of efficiency of spectrum usage.

Despite the fact that the calculated occupancy for some frequency bands is quite low (e.g., the GSM 900 and 1800 uplink bands), CR devices intended to operate in these frequency areas should be properly designed in order to avoid interference with licensed systems.

In order to increase the relevance of the obtained data, measurements in wider frequency bands (up to 6 GHz) are intended in the near future.

Further measurements will be necessary in case of the 470-766 MHz frequency band once the analog TV broadcasting will be completely replaced by the digital one.

ACKNOWLEDGMENT

This work was supported by the Romanian Ministry of Informational Society through the project 106/2011 "Evolution, implementation and transition methods of DVB terrestrial broadcasting using efficiently the radio frequencies spectrum", by Romanian Authority of Scientific Research in the framework of PNCDI 2 "Partnership" through the 20/2012 SaRaT-IWSN project, by the Romanian UEFISCSU PN-2 RU-TE Project no. 18/12.08.2010 and by Romanian contract POSDRU/89/1.5/S/62557.

REFERENCES

[1] J. Mitola III, "Cognitive Radio for Flexible Mobile Multimedia Communications," in Proc. IEEE International Workshop on Mobile Multimedia Communications, 1999, pp. 3-10.

[2] M. A. McHenry, P. A. Tenhula, D. McCloskey, D. A. Roberson, and C. S. Hood, "Chicago spectrum occupancy measurements & analysis and a long-term studies proposal," in Proc. of Workshop on Technology and Policy for Accessing Spectrum (TAPAS), Boston, MA, USA, August 2006

[3] S. D. Jones, E. Jung, X. Liu, N. Merheb, and I.-J. Wang, "Characterization of spectrum activities in the U.S. public safety band for opportunistic spectrum access," in Proc. 2nd IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN 2007), Apr 2007, pp. 137–146.

[4] R. Bacchus, T. Taher, K. Zdunek, and D. Roberson, "Spectrum Utilization Study in Support of Dynamic Spectrum Access for Public Safety," 2010 IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN 2010), Singapore, April 2010.

[5] M. H. Islam et al., "Spectrum Survey in Singapore: Occupancy Measurements and Analyses," in Proc. 3rd International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom 2008), May 2008, pp. 1–7.

[6] M. Wellens, J. Wu, and P. Mähönen, "Evaluation of spectrum occupancy in indoor and outdoor scenario in the context of cognitive radio," in Proc. Second International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrowCom 2007), Aug 2007, p. 8.

[7] R. I. C. Chiang, G. B. Rowe, and K. W. Sowerby, "A quantitative analysis of spectral occupancy measurements for cognitive radio," in Proc. IEEE 65th Vehicular Technology Conference (VTC 2007 Spring), Apr 2007, pp. 3016–3020.

[8] M. Lopez-Benitez, A. Umbert, and F. Casadevall, "Evaluation of Spectrum Occpancy in Spain for Cognitive Radio Applications," in Proc. IEEE 69th Vehicular Technology Conference (VTC 2009 Spring), Barcelona, April 2009.

[9] M. Fadda, M. Murroni, V. Popescu, P. Angueira, J. Morgade, and M. Sanchez, "Hidden node margin and man-made noise measurements in the UHF broadcasting bands," in 2012 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), June 2012, pp.1-5.

[10] M, Lopez-Benitez and F. Casadevall, "Time-dimension models of spectrum usage for the analysis, design and simulation of cognitive radio networks," IEEE Trans. on Vehicular Technology, to be published.

[11] A. Martian, I. Marcu, and I. Marghescu, "Spectrum Occupancy in an Urban Environment: A Cognitive Radio Approach," 2010 Sixth Advanced International Conference on Telecommunications (AICT), May 2010, pp.25-29.

[12] M. López-Benítez and F. Casadevall, "Improved energy detection spectrum sensing for cognitive radio," in IET Communications (The IET), Special Issue on Cognitive Communications, May 2012, vol. 6, no. 8, pp. 785-796.

[13] Z. Tian and G. Giannakis, "A Wavelet Approach to Wideband Spectrum Sensing for Cognitive Radios," în IEEE 1st Int. Conf. on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM), 2006, pp. 1-5.

[14] Y.Zeng and Y.-C. Liang, "Maximum-Minimum Eigenvalue Detection for Cognitive Radio," IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications PIMRC 2007, September 2007, pp.1-5.

[15] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," IEEE Communications Surveys & Tutorials, First Quarter 2009, vol.11, no.1, pp.116-130.

[16] ANCOM, "TNABF", available online at: http://www.ancom.org.ro/uploads/links_files/TNABF_2009+modif-2010_2011.pdf, April 2013.

 ISBN: 978-1-61208-279-0

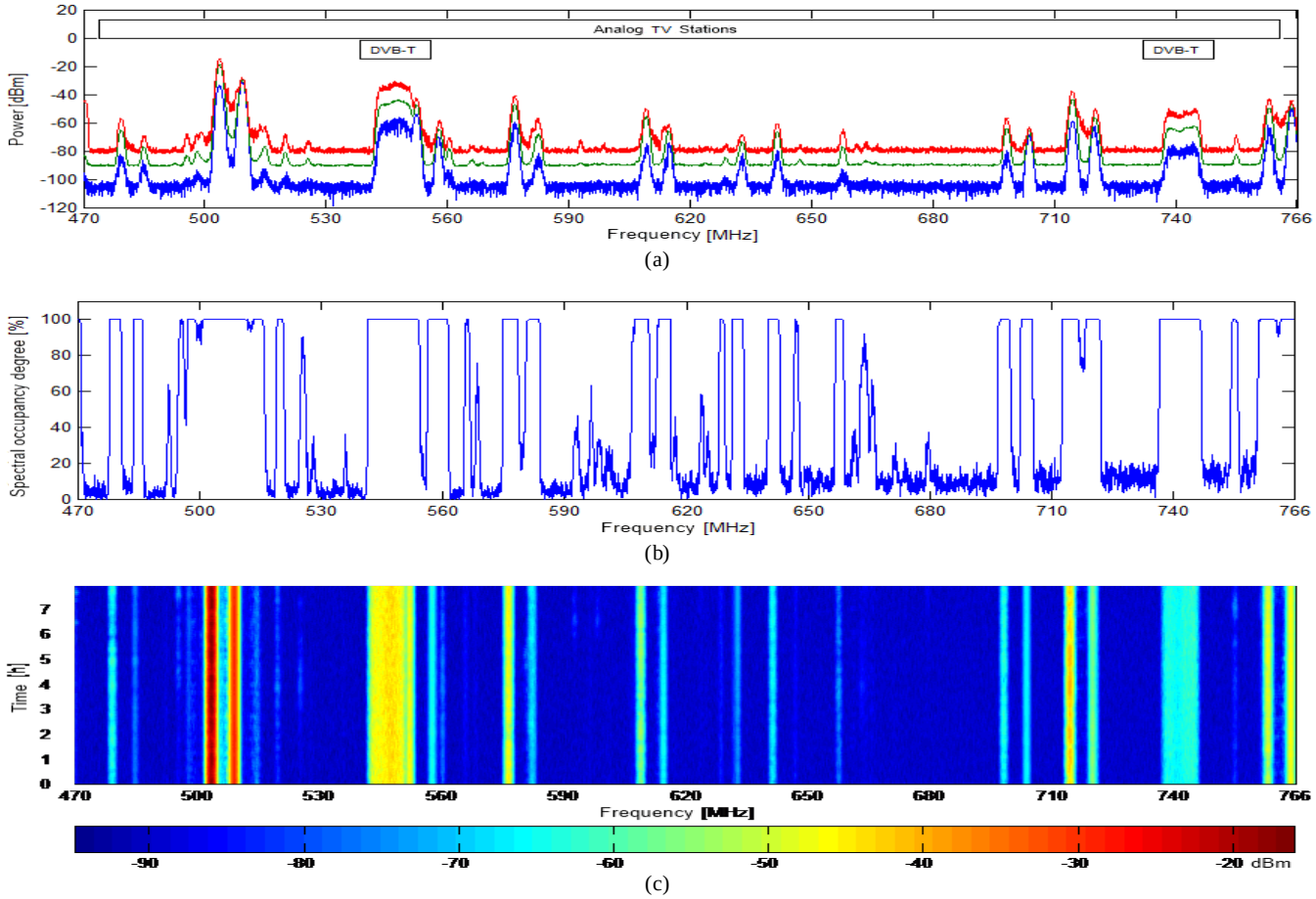


Figure 2. Spectral occupancy for the 470 MHz - 766 MHz frequency band using different metrics (urban environment, 14:00-22:00, working day) (a) Mean, minimum and maximum signal values (b) Occupancy duty cycle (c) 8 hour histogram

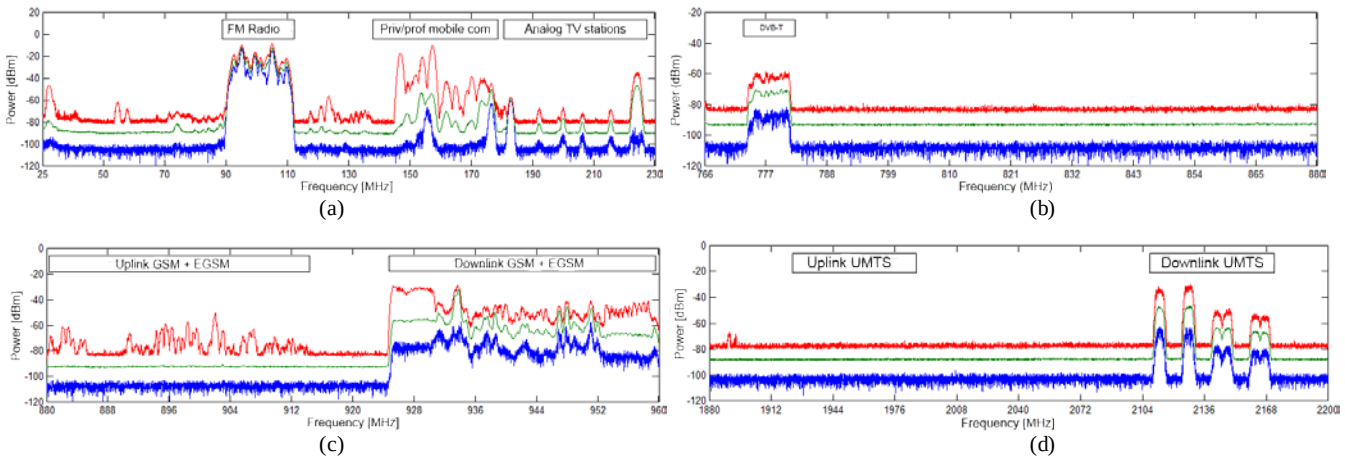


Figure 3. Mean, minimum and maximum signal (urban environment, 14:00-22:00, working day) for the frequency bands (a) 25-230 MHz (b) 766-880 MHz (c) 880-960 MHz (d) 1880-2200 MHz

 ISBN: 978-1-61208-279-0

Security Issues and Threats in Cognitive Radio Networks

Yenumula B. Reddy
Department of Computer Science, Grambling State University
Grambling, Louisiana, USA
ybreddy@gram.edu

***Abstract*—Cognitive radio technology is the vision of pervasive wireless communications that improves the spectrum utilization and offers many social and individual benefits. The objective of the cognitive radio network technology is to utilize the unutilized spectrum by primary users and fulfill the secondary users' demands irrespective of time and location (any time and any place). Due to their flexibility, the cognitive radio networks are vulnerable for numerous threats and security problems that will affect the performance of the network. Little attention was given to security aspects in cognitive radio networks that include spectrum sensing (sensing primary user), attacks that threaten the network at various layers and adversary effects on performance due to the security threats. In this survey, we discuss the cognitive radio networks, problems involved in sensing and management, attacks on cognitive radio networks, attacks on various network layers, threats on cognitive radio networks, and the current security and privacy solutions available. Further, we illustrate the need of careful engineering with security checks while designing the cognitive radio networks.**

Keywords: Cognitive Networks; security; threats; frequency sensing; spectrum mobility; spectrum holes.

I. INTRODUCTION

The novel approach of cognitive radio (CR) in wireless communications was coined by Joseph Mitola III in 1998, in a seminar at Royal Institute of Technology, Stockholm. The work was later published by Mitola and Maguire [1] in IEEE personal communications. The aim was to provide appropriate intelligence (an intelligent agent) to portable devices (for example – personal digital assistants) so that they fulfill the common user communication needs [2]. The intelligent agent in the device detects available channels in the wireless spectrum band and automatically changes its parameters (transmission or reception) to meet user needs. The process of detecting unused or available channels from the wireless spectrum band at any place is called dynamic spectrum access (DSA). The DSA and DSM (dynamic spectrum management) concepts are derived from the principles of artificial intelligence, machine learning, and cross-layer optimization. The game theory applications are examples of DSM that improves the performance of cognitive radio networks (CRN).

The main functions of cognitive radios are spectrum sensing, spectrum management, spectrum mobility and spectrum sharing. The main goal of the cognitive radio is to detect the white spaces (unused spectrum or spectrum holes) in the primary spectrum and efficient use of that detected unused spectrum without harming the primary user. Detection of a transmitted signal can be done by using one or more of techniques including matched filter, energy detection, cyclostationary feature detection, cooperative detection (sensing spectrum with the cooperative effort of multiple cognitive radios), and interference based detection method. The spectrum management (analysis and decision) includes the selection of the best spectrum suitable to the cognitive users. Spectrum mobility is the process of allocating the best possible spectrum during mobility of cognitive user. Finally, the spectrum sharing is a fair scheduling method in spectrum usage.

Today more than 5 billion devices are in use, expected to increase 10 billion by 2017 and approximately 100 billion by 2025. This number includes high-end handsets, tablets, and laptops on mobile networks. These devices generate serious traffic on the communications. The anticipated demand in communications forces to incorporate compact devices, new features, and more battery life. Future cognitive radios offer the new technology with nanotechnology featured compact devices. Building flexible cognitive radio technology with large-scale deployment of cognitive radio networks is a complex task. The features include smart antennas, new hardware (Nano components incorporated) with software defined radio, spectrum sensing, spectrum measurement, medium access control, routing, self-organizing, adaptive control mechanisms, learning, policy definition and monitoring. Developing and introducing new technology requires appropriate security measurements and policies. Therefore, security at each step of cognitive wireless networks is a challenging job.

Berkeley wireless research center [3] shows that 2GHz to 10GHz spectrum is underutilized. To utilize the underutilized spectrum, the cognitive radio must detect the presence of primary signal (PS), and use that spectrum without interfering with primary signal. Security involves in misdetection and false detection of the primary user. False detection is that the primary signal presence is recorded when the signal was absent (falsely detecting the primary user). Further, false detection includes that a malicious user pretends as the primary user (PU) by sending a strong signal to other cognitive users. Misdetection is the presence of the primary user which is

 ISBN: 978-1-61208-279-0

not detected by the cognitive user through matched filter, energy detection, and cyclostationary feature detection. To eliminate such false detection and misdetection of the primary user, a spectrum sensing mechanism must be created to enhance the trust worthiness of the primary signal detection.

In wireless networks, hacking and malicious attacks are inevitable. Further, security threats are unavoidable, and incorporating security facilities are challenging in cognitive radio networks due to its nature of openness. Therefore, more care and research need to be done to provide security mechanisms in cognitive radio networks. Better security mechanisms ensure the trustworthiness of the spectrum sensing. The detection problems arise when operating in a hostile environment. In a hostile environment, it is possible to mimic the incumbent signal characteristics and pretend (emulate the primary user characteristics) as the primary user. In such cases, integrating legitimate transmitters for primary and secondary users in spectrum sensing will improve the trustworthiness of the detection process. Further, embedded signature in PU or interactive protocol between an incumbent transmitter and verifier cannot be used due to FCC's document requirements [4].

In conventional radio technology, signals emitted by wireless devices were predictable (approved by FCC). Since the creation of a wide range of authorized and unauthorized signals are possible using low-cost consumer devices, it is relatively easy to create denial-of-service (DoS) attacks that can affect the critical applications such as traffic control or health care. Therefore, future FCC regulations need to be aware of these DoS attacks [6]. The issues to be considered in cognitive radios are:

- Type of security attacks
- Security implication in implementing software defined radio
- Trusted cognitive radio design with security concerns
- Requirement of authentication protocols in cognitive radio networks
- Ensuring the functionality of security mechanisms and
- Requirement in encryption mechanisms

This survey report focused on the above six problems. Further, the survey identifies and presents the new capabilities to defend against intrusions and denial of service attacks. More work is needed to develop better security models in spectrum sensing, emulation of denial of service, physical layer security enhancements, geo-location for improved wireless network security, and cooperative methods for isolating the intruders. Since the CRN is in a developmental stage, it is an opportunity to incorporate these security capabilities as part of the implementation.

The remaining part of the paper discusses the related work in Section II, cognitive radio network environment in Section II and threats and attacks in section IV. Section 5 provides the type of attacks on a network layer and counter measures. Section VI concludes the work and future requirements in CRN security

II. RELATED WORK

Most of the survey papers in cognitive radio networks (CRN) discussed the security problems in specific aspects of the network. The surveys on CRN show the state of the art research in specific or few general aspects. Fragkiadakis et al [7] discussed the security threats and detection techniques in CRN. The paper includes the challenges that cognitive radios and cognitive radio networks along with the current state-of-the-art to detect the corresponding attacks.

Wassim et al [3] discussed the security attacks along with the mitigation techniques in CRN. This paper provides the category of attacks at MAC layer, data-link layer, and transport layer. Further, they discussed the jamming attack, false detection of PU, and objective function attack which are common in most of the surveys. Their evaluation shows that the combination of the counter measures will produce better security.

The survey paper by Parvin et al [8] addresses the CRN architecture and security issues. The spectrum mobility threats, jamming counter measures, spectrum sensing challenges, and attacks on protocol layers are outlined in this survey. Further, attacks on protocol layer are listed in this survey paper. Leon et al discussed the attacks on cognitive radios including PU emulation attacks, objective function attacks, common control data attack, lion attack, false feedback attack, jamming countermeasures, and vulnerabilities inherent to those systems [9]. They further discussed mitigation of lion attack based on periodic PS emulation attacks. The document is the over view of some of the attacks on cognitive networks. Clancy et al [10] discussed the threats and mitigation of security in cognitive radio Networks. They outlined the policy radio threats, learning radio threats, and self-propagating behavior. Various classes of attacks including dynamic spectrum access (DSA) attacks, objective function attacks, and malicious attacks are part of this document. The authors felt that the earning the trust in cognitive radio networks is extremely important.

Newman and Clancy [11] discussed the security threats in signal classifiers. They discussed the signal classifier model, threat analysis, and threats on feature extractions. They claimed that signal classification algorithm opens a new area of security research related dynamic spectrum access and signal classification. They used the signal classification algorithm to distinguish the primary user (PU) and secondary user signals. Chen et al [12] designed a defense scheme to identify the malicious users by estimating location information and observing the signal strength. Spectrum sensing is also discussed by Chen et al in [13, 14]. In [13], the authors discussed the primary user emulation problem and demonstrated the disruptive effects in cognitive radio networks. They discussed the transmitter

 ISBN: 978-1-61208-279-0

position for detecting the attacks. Further, they demonstrated the effect of the location verification with respect to the attacks. Chen et al [14] discussed the distributed spectrum sensing and incumbent emulation attacks. The sensing and management attacks are:

- Defending against incumbent emulation attacks
- Spectrum sensing data falsification attacks
- Defending against spectrum sensing data falsification (SSDF) attacks

Primary user emulation attack is one of the common security threats in CRN. Chen et al [15] proposed a transmitter verification scheme called LocDef (localization based defense) that verifies the received signal based on location and characteristics. They concluded that the signal disruptive process will be eliminated by incorporating the LocDef process into spectrum sensing processes. They showed through simulations that LocDef scheme is an effective program and can be employed in a hostile environment.

Common control channel security is vital in cognitive radio networks. Safdar and O'Neill [16] discussed the common control channel security for cooperatively communicating cognitive radio nodes. The authors presented an algorithm and demonstrated that low cost hash or message authentication code algorithm achieves information integrity.

The key challenges and evaluation approaches in CRN were presented in [17]. The paper discusses the current security posture of emerging IEEE 802.22 cognitive radio standard and identifies the potential vulnerabilities along with potential mitigation approaches. The features of cognitive radio from the perspective of an attacker were briefly presented. The author identifies that the CR must incorporate the ability to authenticate the local observations in perceived environments, strong collaboration of CR elements related to security, validity of observations between CR elements, and have self-analysis behavior. He further noted that security in CRN is a multi-disciplinary problem.

Implementation issues of spectrum sensing in cognitive radios were discussed by Cabric et al in 2004 [18]. The authors identified cyclostationary feature detection has more advantage among matched filtering and energy detection due to its ability in differentiate modulated signals, interference, and low signal noise ratio. The energy detection technique to detect the primary signal became the central issue of security threats and work on security was concentrated at later years in primary signal emulation analysis. Chen et al [12-15] used various techniques including LocDef for primary signal emulation to eliminate the false detection and misdetection.

III. COGNITIVE RADIO NETWORK ENVIRONMENT AND SECURITY

The increase in communication requirements, difficulties to meet the emergency communication connections and more effective communication services lead the idea of introduction of software defined radio. Cognitive radio is the improvement of software defined radio coined by Mitola [1, 2]. If the demand for spectrum increases continuously, additional radio resources are required to meet customer requirements. Further, we need an agent component that is intelligent enough to adjust transmission parameters with respect to location, environment, and serve the customer needs. This intelligent resource called cognitive radio that operates effectively and use the unused bands (spectrum holes) without disturbing the licensed user. The definitions from National Telecommunications and Intelligent Agency (NTIA), the position statement of IEEE-USA Board of Directors in 2003, and Scientific American conclude that 'cognitive radio is a smart radio that has the ability to sense external environment, learn from history, adjust its parameters to the current state and take intelligent decisions' [19-22]. The statements conclude that CR adapts to the current environment, reasons, learns, collaborates with other radios, and support future decisions. Further, the CRN nodes sense the current radio frequency spectrum environment, contains policy and configuration databases, have self-configuration, mission-oriented configuration, adaptive nature, distributed collaboration and security (authenticate, authorize and protect) of customers.

Due to the nature of CRN, security became a problem at every step (Spectrum Sensing, Spectrum sharing, Location Identification, etc.) of its functionality. The security problems will occur in different ways. For example:

- False detection (sensing) and misdetection of primary signal may happen due to denial of service or malicious user pretends as the primary signal.
- Environment could be controlled by a malicious user.
- An attacker could prevent the cognitive user from using available spectrum (primary signal sensing mechanism).
- An attacker could access the data unauthorized way or modify/inject the false data (integrity of data is required).

Therefore, we need to find the potential threats, potential attacks, likelihood of these threats and attacks, and potential consequences of these attacks. After finding these security risks, we will specify the basic security services such as confidentiality, privacy, and authentication similar to wireless networks. Little attention was given to location privacy threats which are a unique challenge in CRNs. Further, the encryption and mutual authentication techniques will help the data confidentiality.

Figure 1 shows the general architecture of cognitive radio. The security problems are in location identification, the cognitive radios contesting for free spectrum, spectrum sensing, spectrum analysis, and spectrum management. The external threats include hacking the information, incorporating the malicious nodes, corrupting the information at any level shown in the figure 1. Further, the security at the protocol level that interact various layers of

 ISBN: 978-1-61208-279-0

the network is an essential issue to discuss in the study of securing cognitive networks.

Once the spectrum holes are detected, the available spectrum will be allocated as soon as possible. Therefore, cognitive radios are competent among themselves at each node to gain spectrum access. Due to these reasons, the design of CRs requires security and threat procedures. Further, the CRs are vulnerable to the threats and attacks while detecting the primary signal due to their localization and adaptive nature. The malicious attackers are common in wireless networks as well as CRs. The general requirements of security in CRs are discussed in [7-20].

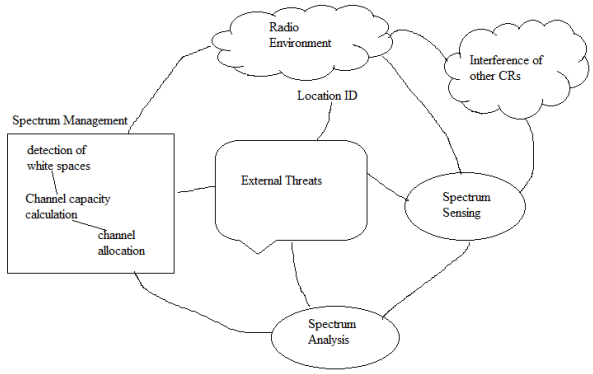


Figure 1. General architecture of cognitive radio

The security requirements in CRNs vary from location to location. Along with security requirements, we will discuss the current state of algorithms for protection in cognitive radio networks and security provisions through IEEE 802.22. The basic security requirements in CRN are confidentiality, integrity and availability. The other security issues include:

- Sensing and emulation of primary signal (detecting and verifying the signal).
- Spectrum management (detecting, verifying channel capacity, and allocating appropriate channel to cognitive user).
- Checking interference level, signal strength, and energy detection.
- Secure communication.

The motivations vary depending upon the attacker. The selfish nature of a cognitive user projects he/she as the primary user to use the spectrum with higher priority. They modify the spectrum sensing parameters for selfish advantage. The selfish user can prevent other users from using the spectrum by jamming or with DoS. The DoS can be created using various authorized and unauthorized waveforms with a low-cost consumer device. The selfish users can be controlled through access permissions and authentication. Further, by using channel sensing algorithms we can control the cognitive users from interference. We will discuss various threats, attacks and the counter measures in Section 4.

To minimize these selfish attacks it is necessary to identify the type of DoS attacks on CRN and possible hardware improvements, improve the weaknesses of Internet to implement software-based radio, trusted CR to address the security issues, and develop the algorithms and mechanisms to address the cooperative methods to detect and isolate intruders.

IV. THREATS AND ATTACKS IN COGNITIVE RADIO NETWORKS

Threat is a constant danger through persons, objects, or any resources where as an attack is an act of or event that exploits the vulnerability. The policies, learning mechanisms, and self-propagation in cognitive radio architecture prevents the threats (cannot escape the threats). In CR, a threat can happen while sensing of information (due to involvement of a malicious user). This information will then feed for learning and decision making. The results produced will lead to inappropriate decisions (unacceptable decisions) due to a malicious user injected the faults. The threat analysis in unsupervised learning and signal sensing was discussed by Clancy et al [10, 11].

Attacks on spectrum management were briefly explained by Parvin et al [8] and Mathur and Subbalakshmi [23, 24]. They suggested strong encryption mechanism is required at physical and MAC layer level. The attacks were classified depending upon the protocol layers. The Table 1 provides the attack types, network layers involved, and reason for attacks.

TABLE 1. ATTACK TYPES, LAYER INVOLVED AND REASON FOR ATTACKS

Attack type	Network Layer	Reason	Countermeasures
Primary and secondary user Jamming	Physical	Lack of knowledge about location and unclear access rights to cognitive user	• Location Consistency Checks • Compare signal strength and noise level
Primary signal sensing	Phycal	Low level primary signal will be missed	• Energy-based sensing • Waveform-based sensing • Cooperative detection of PU
Overlapping secondary users	Physical	Location based. Hard to prevent	• Use game models and Nash equilibrium techniques to detect transmission power of SUs
SUs unauthorized gain in bandwidth by pretends as primary user or False feedback	MAC	Malicious SU tweaks with higher power bandwidth, and feed false information to gain signal	• Trust management on secondary users for resource hungry and collaborative trust • Management of systems objective function by controlling the radio parameters
Increase interference by malicious node	Network	Compromising with malicious node	• Appropriate local spectrum sensing controller • Eliminating internal hidden parasite nodes
Ripple effect	Network	False information about spectrum assignment	• Continuous trust management process on SUs

 ISBN: 978-1-61208-279-0

Key duplication	Transport	Breaks the cypher system	• Reinvestigate the protocol activity in the context of sessions • Use secure protocols with robust distribution of key management
Jelly fish	Transport	Effect on throughput	• Trust of node by verifying the packet loss

Cross-layer attacks are possible in CRN. There is a need to be given individual attention for such attacks. Jamming on routing information happens due to lack of common control channels. Traffic analysis attack on data privacy and location privacy will be avoided by authentication and controlling the access rights of cognitive user.

The defense mechanisms were discussed in [10-20]. The other attacks include false feedback of information from one group of cognitive users to mislead the different group of cognitive users. This consequence ends to mislead the detection of primary signal. Network Endo-Parasite (NEP) attack avoids the selection of the right channel by the other cognitive users. The NEP attack is played by a different group of cognitive users. The objective function attack controls a large number of radio parameters. According to Clancy and Georgen [10] secure communication with low or high power has provided the weights. The channel gain depends upon the weight rate. The dishonest users will mislead the other users to gain access. Further, they mislead the honest user to misdetection of the primary signal with the introduction of extra noise.

V. CHALLENGES AND DEFENDING MALICIOUS ATTACKS

Spectrum sensing, Spectrum management, spectrum sharing, and spectrum mobility are some of the challenges in CRN security. Ensuring trust worthy spectrum sensing is one of the essential mechanisms in CRs. The primary signal analysis is suggested in the current survey. Trust on spectrum sensing happens if the primary signal is emulated and recognized correctly. For example, a malicious user or hacker can interpret the primary user signal and occupy the spectrum for selfish use. The attack can be detected through transmitter verification procedures and location verification procedures. Further, a cognitive user simulates the primary user for personal gains. That means, a cognitive user crosses its user access limits. These activities can be controlled using the various privacy procedures and access limits. This problem can be fixed using a honey pot database to mislead the malicious user.

The primary signal cannot be detected because of interferences at location devices. Primary user signal detection gets difficult if it uses the spread spectrum signaling or altering the parameters by a malicious user. These problems can be eliminated using the cloud application. Further, the cloud application to eliminate the hidden terminal problem was discussed in [25, 26]. The solution for interference problem was proposed in [27] and spectrum sensing can be detected efficiently through multiple users in a cooperative manner. Once the free spectrum is detected, the best available band will be detected using local observations and statistical information.

The common control problem involves the exchange of security keys between the nodes. The authentication among the nodes provides confidentiality and integrity of the transactions. This technique provides the security and the hidden terminal problem still remain. The jamming problem, hidden terminal problem, exchange of keys between the nodes and malicious user acts can be eliminated by using the cloud application. The security to cloud still remains an open problem.

Malicious activity can be from outside or among the cognitive users. Detecting the malicious activity among the cognitive users can be done using the intruder detection procedures and incorporating honey pot database. Further, cross-layer technique with appropriate defense mechanism in communication protocol will help the attacks on upper layers. Incorporating the cryptographic techniques or digital signature based primary signal identification may help in distinguishing the malicious users. More work is required in this direction.

Spectrum mobility involves common control channel, operating frequency range, and location information. It requires the current location of the primary user and operating range so that the secondary user can vacate the occupied spectrum as soon as PU enters. Spectrum mobility depends upon the primary user entry and secondary user relocation. The cloud application will solve many attacks and hidden terminal problems in cognitive networks similar to problems like sudden entry of the primary user [26].

VI. CONCLUSIONS AND FUTURE WORK

The literature shows that the spectrum management sachems are lack of formal security models. The conventional authentication models for wireless security need to be modified to CRNs. The cooperative sensing models improves the sensing capability with overhead of jamming. In multi-user environment, the attackers have opportunities for malicious activities. Intrusion detection models are required in such situations. Cryptographic techniques are useful with the additional burden of computing time. The trust models are more appropriate than to cryptographic techniques due to their simplicity and computational efficiency. Cloud application helps to eliminate hidden terminal problem. But, cloud security is another open problem in CRNs.

The overview of the CRN shows that security is an essential at all levels (sensing, location, and management). Security mechanisms through protocols at different layers were discussed in this paper. The study shows that

 ISBN: 978-1-61208-279-0

implementation at the protocol level is very important at each network layer (Physical, MAC, Network, and transport). Few authors stressed the need of security models at cross-layer design. The current research shows that future security on cross-layer design will get special attention.

Finally, we conclude that threat proof mechanism is difficult and impossible. The threat detection mechanisms can be developed for cognitive radio networks in the same lines of intrusion detection mechanisms. The threat detection and protection of information are serious issues in wireless networks as well as cognitive radio networks. It is recommended that threat detection mechanisms must be developed and incorporated as part of the design as and when need arises.

ACKNOWLEDGEMENTS

The research work was supported by the Minority Leaders Program through contract number GRAM 11-S567-0017-02-C2. The author wishes to express appreciation to Dr. Connie Walton, Provost and Vice President, Academic Affairs, Grambling State University for her continuous support.

REFERENCES

[1] J. Mitola and G. Q. Maguire, "Cognitive Radio: Making software radios more personal", IEEE personal Communications, 1989, vol. 6, no. 4, pp. 13-18.

[2] J. Mitola, "Cognitive Radio – An Integrated Agent Architecture for Software Defined Radio", Ph.D. Dissertation, Royal Institute of Technology, Kista, Sweden, May 8, 2000, ISSN: 14035286, 313 pages..

[3] E. Wassim, S. Haidar and G. Mohsen, "Survey of Security Issues in Cognitive Radio Networks", Journal of Internet Technology, 2011, vol. 12 No. 2, pp. 181-198.

[4] FCC, "Notice for Proposed Rulemaking (NPRM 03-322): Facilitating Opportunities for flexible, Efficient, and Reliable Spectrum Use Employing Cognitive Radio Technologies," ET Docket, No. 03-108, 2003.

[5] D. Sicker and R. Dhillon "Security of Cognitive Radio Networks (Synthesis Lectures on Communications)", Morgan & Claypool Publishers (January 30, 2013), ISBN-13: 978-1608451005.

[6] Federal Communication Commission, "Unlicensed operation in the TV broadcast bands and additional spectrum for unlicensed devices below 900 MHz in the 3GHz band," ET Docket, No. 04-186, May 2004.

[7] A. G. Fragkiadakis, E. Z. Tragos and I. G. Askoxylakis, "A Survey on Security Threats and Detection Techniques in Cognitive Radio Networks", IEEE Communications Surveys & Tutorials, 2013, vol. 15 , issue: 1 , pp. 428 -445.

[8] S. Parvin, F. K. Hussain, O. K. Hussain, S. Han, B. Tian, and E. Chang, "Cognitive radio network security: A survey", Journal of Network and Computer Applications, 2012, vol. 35, pp. 1691–1708.

[9] O. León, J. Hernández-Serrano, and M. Soriano, "Securing cognitive radio networks", Int. Jr. of Communication Systems, 2010, vol. 23, Issue 5, pp. 633-652.

[10] T. C. Clancy and N. Goergen, "Security in Cognitive Radio Networks: Threats and Mitigation", CrownCom 2008, 3rd International Conference on Cognitive Radio Oriented Wireless Networks and Communications, 2008, pp. 1 – 8.

[11] T. R. Newman and T. C. Clancy, "Security threats to cognitive radio signal classifiers", Proceedings of the Virginia tech wireless personal communications symposium, 2009, pp. 1-9.

[12] K. C. Chen, Y. J. Peng, N. Prasad, N., Liang, Y. C. and S. Sun "Cognitive radio network architecture: part I. general structure", 2nd international conference on ubiquitous information management and communication, 2008, CRs. pp.114–119.

[13] R. Chen and J. Park, "Ensuring Trustworthy Spectrum Sensing in Cognitive Radio Networks", 1st IEEE workshop on Networking Technologies for Software Defined Radio Networks,(SDR '06), 2006, pp. 110 – 119.

[14] R. Chen, J. Park, Y. T. Hou and J. Reed, " Toward Secure Distributed Spectrum Sensing in Cognitive Radio Networks", IEEE Communications Magazine, 2008,vol. 46, issue. 4, pp. 50-55.

[15] R. Chen, J. Park and J. H. Reed, "Defense against Primary user Emulation Attacks in Cognitive Radio Networks", IEEE Journal on selected areas in communications, vol. 26, no.1, 2008, pp. 25-37.

[16] G. A. Safdar and M. O'Neill, "Common Control Channel Security Framework for Cognitive Radio Networks", 69th IEEE Vehicular Technology Conference, 2009, pp. 1-5.

[17] J. L. Burbank, "Security in Cognitive Radio Networks: The Required Evoluation in Approaches to Wireless Network Security", CrownCom 2008, pp. 1-7.

[18] D. Cabric, S. M. Mishra and R. W. Brodersen, "Implementation Issues in Spectrum Sensing for Cognitive Radios", 38th Asilomar Conference on Signals, Systems and Computers, 2004, pp. 772-776.

[19] "IEEE 802 Tutorial: Cognitive Radio", Scott Seidel, Raytheon, presented at IEEE 802 Plenary, 18 July 2005.

[20] T. R. Shields, "SDR Update," Global Standards Collaboration, Sophia Antipolis, France, Powerpoint Presentation GSC10_grsc3 (05)20, 2005.

[21] R. W. Thomas, L. A. DaSilva and A. B. MacKenzie, "Cognitive networks," IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, November, 2005, pp. 352-360.

[22] FDR Forum, Cognitive Radio Definitions and Nomenclature, Approved Document SDRF-06-P-0009-V1.0.0, 10 September 2008.

[23] C. N. Mathur and K. P. Subbalakshmi, "Security issues in cognitive radio networks", Cognitive networks: towards self-aware networks. John Wiley and Sons, Ltd; 2007.

[24] C. N. Mathur and K. P. Subbalakshmi, "Digital signatures for centralized DSA networks", 4th IEEE conf. on consumer comm. and networking, 2007, pp. 1037–1041.

[25] Y. B. Reddy, "Solving Hidden Terminal Problem in Cognitive Networks Using Cloud Application", SENSORCOMM 2012, pp. 235-240.

[26] Y. B. Reddy and S. Ellis, "Modeling Cognitive Radio Networks for Efficient Data Transfer Using Cloud Link", ITNG 2013, April 2013, Las Vegas, USA.

[27] M. Shahid and J. Kamruzzama, "Agile spectrum evacuation in cognitive radio networks", IEEE international conference on communications (ICC), 2010, pp. 1–6.

 ISBN: 978-1-61208-279-0

Spectrum Sensing Using Sub-Nyquist Rate Sampling

Zahid Saleem, Samir Al-Ghadhban
Department of Electrical Engineering
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia
Email: (zahidsaleem, samir) @ kfupm.edu.sa

***Abstract*—Spectrum sensing in wideband regime requires huge amount of samples. The observed frequency spectrum is usually sparse. Compressed sensing technique provides a viable solution to reconstruct the sparse signals. The observed wideband spectrum can be reconstructed using compressive sensing technique. Inherent constraints of the compressed sensing algorithms hinder the flexible implementation of spectrum sensing process. The structure-based Bayesian sparse recovery algorithm is used in this paper to implement spectrum sensing process. Spectrum sensing performed using the Bayesian estimation approach resulted in better performance compared to the results based on compressed sensing technique. Various cases have been discussed considering the amount of information available for the observed frequency band. Spectrum sensing performed using the Bayesian algorithm showed improvement of more than 5 dB in all cases.**

Keywords; cognitive radio; spectrum sensing; compressive sensing; structure-based bayesian sparse recovery algorithm.

I. INTRODUCTION

The ever-increasing high data rate services and new wireless service providers require more frequency spectrum than available. This appetence of more frequency spectrum has raised a concern of spectrum scarcity. The frequency spectrum is a limited natural resource. Measurements have shown that the current spectrum scarcity is a result of under-utilization rather than the unavailability of spectrum. According to Federal communication comission [1], the spectrum utilization varies from 15% to 85% with high variance in time and space. These statistics puts question on the appropriateness of current regulatory authorities. To overcome this problem, Mitola and Maguire [2] introduced the cognitive radio device in 1999. The cognitive radio (CR) provides an adequate solution to the observed concern of spectrum scarcity. The CR avails opportunistic access to the frequency bands that are not used by the licensed users at a particular instance or space [2].

This paper focuses on performing spectrum sensing in wideband regime. The spectrum sensing process is core of the CR system. It enables the CR to scan range of frequencies and utilize any vacant ones. This process has many challenges associated with it. One key problem relates to the sensing of a wideband signal. Perpetually sampling of signal is done at the Nyquist rate. In the wideband regime, this means acquisition of colossal amount of samples and respectively high sensing time. In this paper solution to the wideband spectrum sensing problem is discussed using the sub-Nyquist rate sampling technique.

Over the years, many algorithms (based on the Nyquist rate sampling criteria) have been developed for the spectrum sensing process. Among them are the energy detection based sensing, wave form based sensing, cyclostationary feature based sensing and match filtering based sensing [4]. Sensing a wideband signal using these techniques require large amount of time. The compressive sensing (CS) technique provides reconstruction of the sparse signals sampled at less than the Nyquist rate [5].

Over the last few years some algorithms have been proposed that perform spectrum sensing using the compressive sensing technique. Some of these algorithms are discussed in this paragraph. In 2007, Tian and Giannakis [6] proposed the idea of performing spectrum sensing using the compressive sensing technique. As the observed signal is sparse in frequency domain, its frequency spectrum was reconstructed using the compressive sensing technique. The estimates of various frequency band locations (within the observed spectrum) were generated using the wavelet edge detection technique. The presence or absence of a primary user within each frequency band was determined by observing the corresponding power spectral density (PSD). In 2009, Polo et al. [7] used the analog to information Converter (AIC) instead of the analog to digital converter (ADC) at the receiver. An AIC can be conceptually viewed as an ADC operating at the Nyquist rate followed by the compressive sampling mechanism. In 2009, Chen et al. [9] improved the work proposed in [6]. A multi-branched spectrum sensing structure was proposed. Each branch repeats the same procedure proposed in [6], i.e., reconstructs the frequency spectrum of received signal and calculates the PSD within each band. The results from all branches were combined to generate a final estimate. In 2010, Nassab et al. [8] assumed a fixed number of frequency bands in the observed spectrum. The wideband filters were used to acquire energies from some frequency bands. As the complete energy vector of the observed spectrum is sparse in nature, it was recovered using the compressive sensing technique. In 2010, Sundman et al. [10] modified the

 ISBN: 978-1-61208-279-0

proposed work of [7]. The autocorrelation vector achieved in [7] deals with the wide-sense stationary (WSS) signals only. However, the signal at the output of AIC is non-WSS. The autocorrelation vector was modified in order to deal with the non-WSS signals. They also proposed memory based spectral detection, which resulted in the overall reduction of computational complexity. In 2010, Liu and Wan [11] used the *a priori* knowledge of spectrum distribution and proposed a mixed l_2/l_1 norm de-noising operator. They suggested to attain the primary user frequency band information from the regulatory authorities. The *a priori* knowledge of the band gaps and the block sparsity resulted in better performance when compared to the standard mixed l_2/l_1 norm de-noising operator.

Though compressive sensing algorithms reconstruct the sparse signals with good probability, they do suffer from some deficiencies. They are computationally complex, do not use the structure of the sensing matrix and do not use the *a priori* statistical information about signal support and noise. These algorithms are bottlenecked by the number of observations. Increasing the number of observations leads to the better performance and vice versa. In order to overcome these shortcomings the structure-based Bayesian sparse recovery algorithm (SBBSR) is proposed in [12].

This paper focuses on utilizing the SBBSR algorithm and performing spectrum sensing at the sub-Nyquist rate. The SBBSR algorithm allows flexible implementation in contrast to the compressed sensing based algorithms. The rest of paper is organized as follows. Section II describes the spectrum sensing process performed using the SBBSR algorithm. Section III exploits the flexible implementation of SBBSR algorithm to improve performance. Simulations results are shown and discussed in section IV. Section V provides conclusion to this paper.

II. SPECTRUM SENSING USING THE SBBSR ALGORITHM

The SBBSR algorithm provides reconstruction of sparse signals using the Bayesian estimation approach. While reconstructing the signal it uses the *a priori* statistical and sparsity information and the sensing matrix structure. Assume the sensing timing window is defined as $t \in [0, NT_0]$ (where T_0 represents the Nyquist sampling rate). According to the Nyquist theorem, N samples are required to reconstruct the original signal without aliasing. The sampling process at a digital receiver can be expressed as

$$\boldsymbol{y} = \boldsymbol{\Theta x} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{x}$ represents the $N \times 1$ length sparse vector, $\boldsymbol{\Theta}$ is an $M \times N$ projection matrix (or sensing matrix, which is incoherent with the domain in which $\boldsymbol{x}$ is sparse) and $\boldsymbol{n}$ is the complex additive white Gaussian noise vector $\mathcal{CN}(0, {\sigma^2}_n I)$. The process defined in (1) can be explained as the conversion of a continuous domain signal $\boldsymbol{x} \in \mathcal{C}^N$ into the discrete sequence $\boldsymbol{y} \in \mathcal{C}^M$. In (1) when $M = N$ the Nyquist rate uniform sampling is performed whereas setting $M < N$ performs the reduced rate sampling scheme or the sub-Nyquist rate sampling [6].

The sparse signal $\boldsymbol{x}$ can be modeled as

$$\boldsymbol{x} = \boldsymbol{x}_B \odot \boldsymbol{x}_G \tag{2}$$

where $\odot$ represents dot multiplication between the two vectors, $\boldsymbol{x}_B$ is an independent and identically distributed (i.i.d) Bernoulli random variable and the entries $\boldsymbol{x}_G$ can be drawn from any distribution. This model of $\boldsymbol{x}$ provides a sparse signal. The sparsity information is indulged by the Bernoulli random variable and the amplitudes of these observations are drawn from some other distribution [12].

If the support S of $\boldsymbol{x}$ is known it can be written as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{\Theta x} + \boldsymbol{n} \\ &= \boldsymbol{\Phi \Psi x} + \boldsymbol{n} \\ \boldsymbol{y}|S &= \boldsymbol{\Theta}_S \boldsymbol{x}_S + \boldsymbol{n}_S \end{aligned} \tag{3}$$

$\boldsymbol{\Theta}_S$ is the sub-matrix formed from $\boldsymbol{\Theta}$ containing only those columns represented by S. The maximum a posteriori (MAP) estimate of observed signal $\boldsymbol{x}$ is given as [12]

$$\hat{\boldsymbol{x}}_{MAP} = arg\ max_S\, p(\boldsymbol{y}/S)\, p(S) \tag{4}$$

where $p(S)$ is the probability of a given support. Assuming the signal model of (2), the probability of support can be written as [12]

$$p(S) = p^S (1-p)^{N-S} \tag{5}$$

Now, the problem of calculating MAP narrows down to the calculation of $p(\boldsymbol{y}/S)$. In this paper it is assumed that primary user data has Gaussian distribution, $\boldsymbol{x}|S$ is Gaussian, then $\boldsymbol{y}|S$ will also be Gaussian with zero mean and covariance $\boldsymbol{\Sigma}_S$. Corresponding probability is calculated as [12]

$$p(\boldsymbol{y}/S) = \frac{exp(-\frac{1}{{\sigma_n}^2} \boldsymbol{y}^H {\boldsymbol{\Sigma}_S}^{-1} \boldsymbol{y})}{det(\boldsymbol{\Sigma}_S)} \tag{6}$$

where covariance matrix is given as

$$\boldsymbol{\Sigma}_S = \boldsymbol{I} + \frac{{\sigma_x}^2}{{\sigma_n}^2} \boldsymbol{\Theta}_S {\boldsymbol{\Theta}_S}^H \tag{7}$$

To perform the spectrum sensing process using SBBSR algorithm following steps are opted. These steps are also described in Fig. 1.

1- The sub-Nyquist rate sampled signal $\boldsymbol{y}$ is correlated with the sensing matrix $\boldsymbol{\Theta}$.
2- Based on the correlation result P clusters are made.

 ISBN: 978-1-61208-279-0

3- Let P_c denote the maximum possible support size in a cluster. For each cluster find the likelihoods for all support size starting from $l = 1,2, \ldots P_c$.
4- Within each cluster the MAP estimates of corresponding likelihoods are calculated as explained in (4).
5- Decision regarding presence or absence of the primary user on certain frequency band is made based upon the MAP estimate. The indexes of maximum valued estimates correspond to the occupied locations by a user.

III. EXPLOITING FLEXIBLE IMPLEMENTATION OF THE SBBSR ALGORITHM

The SBBSR algorithm is used to recover the locations where transmission has been done by primary user. Numerous conditions can be imposed to enhance the sensing ability of a CR. These conditions have been discussed in this section and will be used in the simulation part.

- **Case 1**: Considering only signal sparsity as an assumption for spectrum sensing.
- **Case 2**: In addition to sparsity, assuming the observed spectrum consists of fixed (same) length frequency bands.

Consider a scenario in which *a priori* information about the primary user frequency band is available. As proposed in [11], regulatory authorities assign a certain frequency band to a user following the static spectrum allocation scheme. For instance, the bands 1710-1755 MHz and 1805-1850 MHz are allotted to GSM 1800. This also provides a hint that on a certain frequency band the primary users will appear in the form of clusters. For the observed frequency spectrum these details can be gathered *a priori* from the regulatory authority. Here, it is assumed that on a given spectrum all primary users have been assigned known and fixed length bands. One key advantage is the reduction of computational complexity. Earlier calculation of the estimates for various support sizes $l = 1, 2, \ldots P_c$ was required. This resulted in calculation of 2^{P_c} estimates. Now with the length knowledge, the estimates

Begin
Correlate Observation vector $\boldsymbol{y}$ with sensing matrix $\boldsymbol{\Theta}$
Form P semi-orthogonal clusters of length L each around the positions with high correlation values
Process each cluster independently and in each cluster calculate the likelihoods for support of size $l = 1,2, \ldots P_c$
Evaluate MAP estimate
END

Figure 1. Spectrum Sensing Using SBBSR Algorithm

for various support sizes are not required. One estimate is calculated for each cluster.

- **Case 3**: In addition to sparsity, assuming the observed spectrum consists of variable length frequency bands.

Assume that in the observed spectrum, variable length frequency bands are present. The length of these frequency bands is assigned based on some probability distribution function. Assume that this *a priori* length information is also known at the receiver.

IV. SIMULATIONS

The sensing matrix in case of spectrum sensing is a partial inverse discrete Fourier transform (IDFT) matrix and is given as

$$\boldsymbol{\Theta} = \boldsymbol{S_c}^T \boldsymbol{F_N}^{-1} \tag{8}$$

where $\boldsymbol{S_c}$ is the identity matrix of size $N \times M$ and $\boldsymbol{F_N}^{-1}$ is the IDFT matrix of size $N \times N$. In this case the observation vector can be written as

$$\boldsymbol{y} = \boldsymbol{S_c}^T \boldsymbol{F_N}^{-1} \boldsymbol{x} + \boldsymbol{n} \tag{9}$$

$\boldsymbol{n}$ is the complex additive white Gaussian noise vector $\mathcal{CN}(0, \sigma^2{}_n I)$. Here it is assumed that the wideband signal of interest lies in the range of [0,1000]Δ Hz, where Δ is frequency resolution. There are two primary users present in the observed spectrum and are shown in Fig. 2. The observed spectrum is sparse with a sparsity level of 6% and possesses the same structure as described in [6]. This choice of model is helpful in comparing the results of SBBSR algorithm and the approach proposed in [6]. In [6], compressive sensing technique was used for the spectrum sensing process. Frequency spectrum was recovered from the sub-Nyquist rate sampled observations using the l_1 minimization approach. In order to obtain the frequency band edge information the wavelet edge detection technique was applied on the recovered spectrum. The PSD of each frequency band is calculated and decision regarding presence or absence of the primary user is made. In simulation, Gaussian wavelet is used for the edge detection technique.

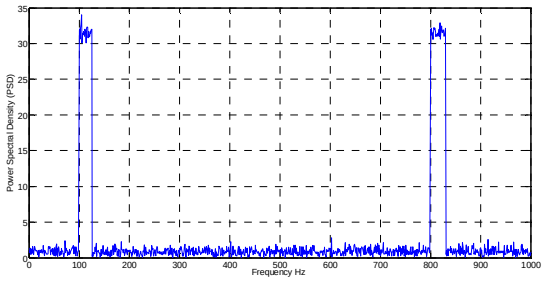


Figure 2. Assumed Wideband Signal-Flat PSD

Aforementioned cases are considered and compared to the compressed sensing approach of [6]. Table I provides the values required by the SBBSR algorithm for these different cases and Table II shows the corresponding working ranges for probability of detection greater than 0.9 for both techniques. Fig. 3 shows the corresponding plots of probability of detection versus signal to noise ratio (SNR).

The spectrum sensing performed using SBBSR algorithm showed better performance than the compressive sensing technique. In all cases, gain of (approximately) more than 5 dB over SNR is observed. The flexible implementation of SBBSR algorithm allowed improvement in the working range of a CR.

V. CONCLUSION

In this paper, the structure-based Bayesian sparse reconstruction algorithm (SBBSR) was used for the spectrum sensing of wideband signals. The SBBSR algorithm provides sub-Nyquist rate sampling solution to the wideband spectrum sensing problem. Spectrum sensing was performed for various cases using both the SBBSR algorithm and the compressed sensing based technique. The results obtained from the SBBSR algorithm showed better performance compared to the other technique. It provided an improvement of more than 5 dB in signal to noise ratio for the observed spectrum. The *a priori* knowledge of the frequency band (whether fixed or variable) helped to achieve better performance. Hence, the SBBSR algorithm improves the performance of spectrum sensing process for the wideband signals and in addition overcomes the shortcomings of compressed sensing technique.

TABLE I. REQUIRED VALUES BY SBBSR ALGORITHM

Cases	Observation Vector Size M	Number of Cluster P	Maximum Support Size P_c	Cluster Size L
1	$\frac{N}{4}$	29	9	9
2	$\frac{N}{4}$	79	1	31
3	$\frac{N}{4}$	79	3	[25 31]

TABLE II. COMPARISON OF CASES FOR PROBABILITY OF DETECTION ≥ 0.9

	Compressed Sensing	Case 1	Case 2	Case 3
SNR ≥	12.95 dB	7.1 dB	7.8 dB	1.8 dB

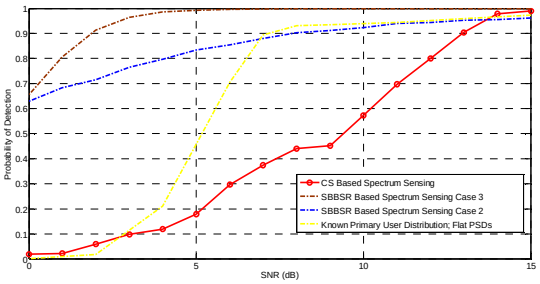


Figure 3. Probability of Detection versus SNR for Case 1

VI. ACKNOWLEDGEMENT

The authors would like to acknowledge the support provided by King Fahd University of Petroleum and Minerals (KFUPM) and King Abdulaziz City for Science and Technology (KACST) through the Science and Technology Unit at KFUPM for funding this work through project number 09-ELE781-4 as part of the National Science, Technology and Innovation Plan.

REFERENCES

[1] Federal Communications Commission – First Report, and order and Further Notice of Proposed Rulemaking, "Unlicensed operation in the TV broadcast bands," FCC 06-156, Oct. 2006.

[2] I. Mitola, J. and J. Maguire, G.Q.,"Cognitive radio: making software radios more personal," IEEE Personal Commun. Mag., vol. 6, no. 4, pp. 13-18, Aug. 1999.

[3] Z. Tian, and G. B. Giannakis, "A Wavelet Approach to Wideband Spectrum Sensing for Cognitive Radios," Proc. of Intl. Conf on CROWNCOM, Mykonos,pp. 1-5, Greece, June 2006.

[4] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications", IEEE Communications Surveys and tutorials, vol. 11 no. 1. First quarter 2009.

[5] R. G. Baraniuk, "Compressive Sensing," Proceeding of IEEE Signal Processing Magazine, pp. 118-124, July 2007.

[6] Z. Tian and G. B. Giannakis, "Compressed sensing for wideband cognitive radios," Proc. Of the International Conference on Acoustics, Speech, and Signal Processing, pp. IV/1357-IV/1360, Apr. 2007.

[7] Y. L. Polo, Y. Wang, A. Padharipande, and G. Leus, "Compressive wide-band spectrum sensing," in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, Taipei, Apr. 2009, pp. 2337-2340.

[8] V. H. Nassab, S. Hassan, and S. Valaee, "Compressive detection for wide-band spectrum sensing," Proc. Of the International Conference on Acoustics, Speech, and Signal Processing, March 2010, pp. 3094-3097.

[9] X. Chen, L. Zhao, and J. Li , "A Modified Spectrum Sensing Method for Wideband Cognitive Radio Based on Compressive Sensing," Fourth International Conference on Communications and Networking in China 2009, pp. 1-5, Aug. 2009.

[10] D. Sundman, S. Chatterjee, and M. Skoglund, "On the use of compressive sampling for wide-band spectrum sensing," in The 10th IEEE International Symposium on Signal Processing and Information Technology, 2010, pp. 354-359.

[11] Y. Liu and Q. Wan, "Compressive Wideband Spectrum Sensing for Fixed Frequency Spectrum Allocation," Science And Technology, p. 21, May 2010.

[12] A.A. Quadeer and T.Y. Al-Naffouri, "Structure-Based Bayesian Sparse Reconstruction," IEEE Transactions on Signal Processing, vol. 60, pp. 6354-6367, July 2012.

 ISBN: 978-1-61208-279-0

An Efficient Image Processing on Sensor Networks

Ben-Shung Chow
Electrical Engineering Department
National Sun Yat-Sen University
Kaohsiung, Taiwan 80424, ROC
bschow@mail.ee.nsysu.edu.tw

***Abstract*—The power consumption rate and bandwidth gain are the two major design criterions considered in the sensor networks, since the sensor nodes in the network are usually with limited capabilities in terms of processing power and bandwidth availability. These two criterions are usually contradictory to each other. However, the efficiencies of power consumption rate and bandwidth gain are achieved simultaneously in the image processing on sensor network proposed in this paper. This is made possible by a technique innovation: image processing in the compressed form. It is noted that in the conventional approach the compressed image must be decompressed first to be processed. This innovation makes the compression and processing a joint design not two separate processes as before.**

Keywords- morphological image processing; quad tree; sensor network

I. INTRODUCTION

Signal processing and data compression should be considered together as a whole to make totally efficient. More specifically, we propose in this paper to have signal processing and data compression suitable for being aware, reasoning and thus convenient for the sensor network [1-2] to make a decision. We accomplish this purpose by making data compressed and signal processed both functionally, i.e., being able to be accessed and processed to a specific interest. More clearly, the data is compressed in the hierarchical data structure processed by morphological image processing. Most importantly, we implement our morphological image processing on the hierarchical data structure [3-4] instead of on the pixel matrix. Therefore, the signal processing and compression are joined together. To our knowledge, this joint processing concept is newly proposed.

The morphology processing has long been applied successfully to industry auto-inspection and medical image processing [5-6]. There are many efforts on hardware implementation to facilitate the morphological processing [7-10]. Neural network consideration was also investigated in [11]. Morphological processing has recently been applied to video coding in sensor network by shape compensation [12], and cognitive sensing [13-14]. The quad tree data structure has been well applied to the field of computer vision such as image segmentation and compression [15-16]. The quad tree with its hierarchical data structure is computationally efficient. The quad tree is based on the principle of recursive decompositions of space. The technique of recursive call has been further developed to extend to the imbedded system [17-18] applications.

The conventional morphological image processing is processed by a pixel vs. pixel basis with the pictures stored in the matrix form. In contrast, the proposed morphological processing in the quad tree data structure can be processed by a block vs. block basis. The pyramid approach can be also classified as fast block basis processing but without the compression consideration due to the pyramid layer of every resolution in matrix form. The proposed approach is novel to achieve fast morphological processing jointly with image compression.

The power consumption rate and bandwidth gain are the two major design criterions considered in the sensor networks, since the sensor nodes in the network are usually with limited capabilities in terms of processing power and bandwidth availability. These two criterions are usually contradictory to each other. However, the efficiencies of power consumption rate and bandwidth gain are achieved simultaneously in the image processing on sensor network proposed in this paper. This is made possible by a technique innovation: image processing in the compressed form. It is noted that in the conventional approach the compressed image must be decompressed first to be processed. This innovation makes the compression and processing a joint design not two separate processes as before.

II. IMPLEMENTATION CONCEPTS WITH WORKING EXAMPLE

It is noted that by the quadtree decomposition, any binary image can be decomposed into black and white square blocks with some fixed size of power of 2. Thus, dilation of the whole image can be accomplished by dilating individual decomposed square blocks. White square blocks do not need dilation, black square blocks take a simple dilation by the tree pattern copy operations. The tree patterns are the dilation results of black square blocks and can be computed as patterns in advance for preparation. It is common for very many dilated block to share the same tree patterns. Therefore, the implementation concepts can be summarized as two strategies: block vs. block processing using pattern method and efficient transformation from sequential access string format to direct access pointer format. A working example is illustrated in Fig. 1.

 ISBN: 978-1-61208-279-0

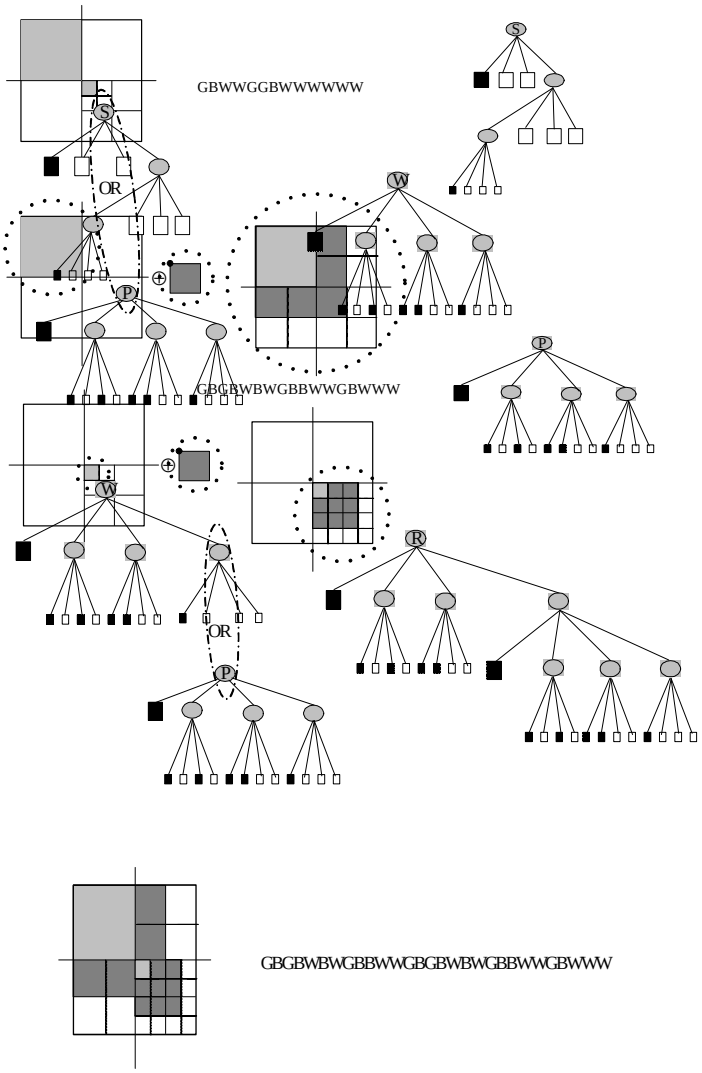


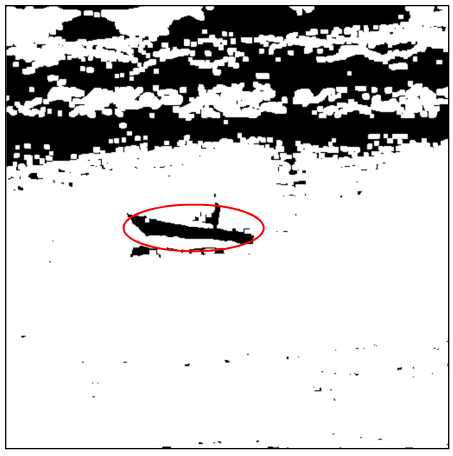
Fig. 1. Working example of dilation by the tree pattern copy operations.

III. EXPERIMENTS ON SENSING APPLICATION

For sensing need, the workout example is a continuation to our research on low bit-rate video coding for sensor network. This continuation extends video coding to target tracking. The goal of this working example is to find object "boat", which is also set for the performance comparison. The boat obtained from the best resolution is regarded as the benchmark as in Fig. 2. The boats obtained from the lower resolutions appear image quality degradation as shown in Fig. 3. Accordingly, the quality degradation can be interpreted as "distortion", which is further numerically defined as the error count from the error image referenced to the benchmark for every resolution. The error images are shown in Fig. 4.

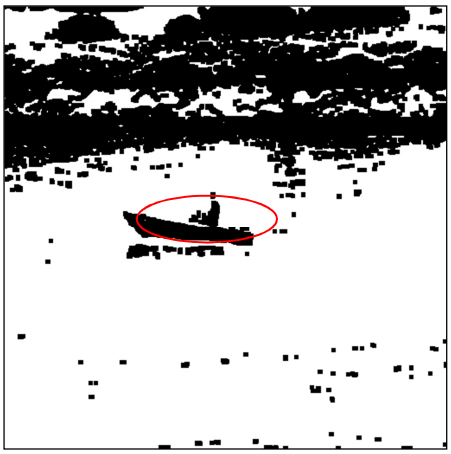
(a)

(b)



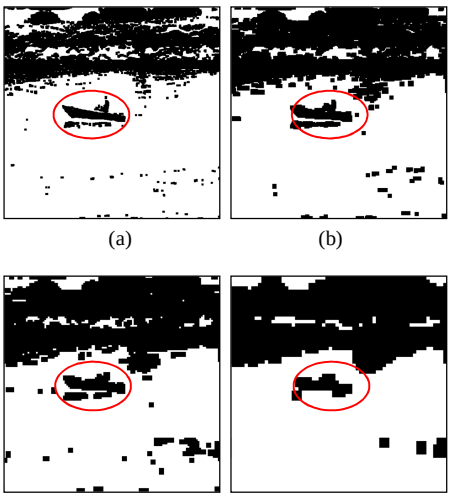
(c)

 ISBN: 978-1-61208-279-0



(d)

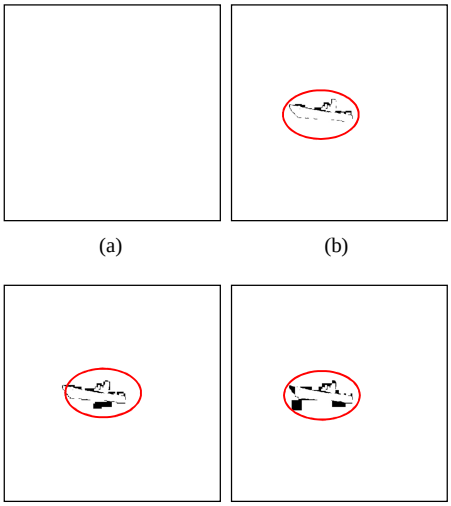
Fig. 2. Intermediate results of working example for target tracking using four resolutions: only the highest resolution is shown in this figure (a) the original 512*512 grey picture (b) the 512*512 binary picture (c) eroding b by structuring element 4*4 square (d) dilating c by structuring element 4*4 square



(a) (b)

(c) (d)

Fig. 3. Final results of working example: eroding, dilating using different resolutions compared to Fig. 5 (but Fig. 5d repeated here as Fig. 6a) (a) the highest resolution: 512*512 original picture with 4x4element (b) the second resolution: 256*256 original picture with 4x4 element (c) the third resolution: 128*128 original picture with 2x2 element (d) the fourth resolution: 64*64 original picture with 2x2 element



(a) (b)

(c) (d)

Fig. 4. Error results of working example: The target error is referenced to the target obtained by the highest resolution in Fig. 6a as the benchmark (a) the first resolution is a blank picture because there is no error at all (b) the second resolution (c) the third resolution (d) the fourth resolution.

As a summary of our experiments, three procedures are applied in our signal processings: (I) binarized quantization, (II) down sampling in resolution (III) shape compensation by morphological operations. It is noted that the error results are compared for different resolutions in procedures II and III.

IV. RATE–DISTORTION PERFORMANCE ANALYSIS

Rate–distortion theory is a major branch of information theory about source coding, which provides the theoretical foundations for lossy data compression; it addresses the problem of determining the smallest compression rate R to encode the source signal without exceeding a given distortion D when reconstructed by the coded data bits. The compression rate corresponds to the cost of data bits and the distortion represents the performance. To apply the rate-distortion theory to our work, we first define the distortion as the error counts from the error images referenced to the target image computed by the best resolution. Secondly, we interpret the rate as the power consumption rate instead of the original data rates.

 ISBN: 978-1-61208-279-0

TABLE I

NODES INFORMATION FOR THE PICTURE "BOAT"

	Gray	Black	White
Level 9	1	0	0
Level 8	4	0	0
Level 7	16	0	0
Level 6	64	0	0
Level 5	247	8	1
Level 4	886	67	35
Level 3	2861	281	402
Level 2	7473	1277	2694
Level 1	13472	6795	9625
Level 0	0	26168	27720

TABLE II
SPEED EFFICIENCY COMPARISON

Methods	Direct Method	Propose Method
Structuring Element	time (msec)	Time(msec)
2 x 2 square	4468	375
4 x 4 square	10764	313
8 x 8 square	31640	406
distributed squares	13483	1172

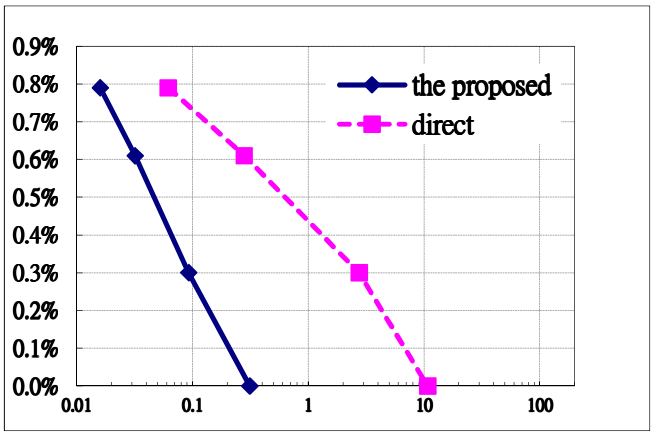


Fig. 5. The rate-distortion performance comparison in consumption power for two methods: the logarithm in the processing (horizontal) axis is also required because of too large deviation for the two processing power spent in the respective methods.

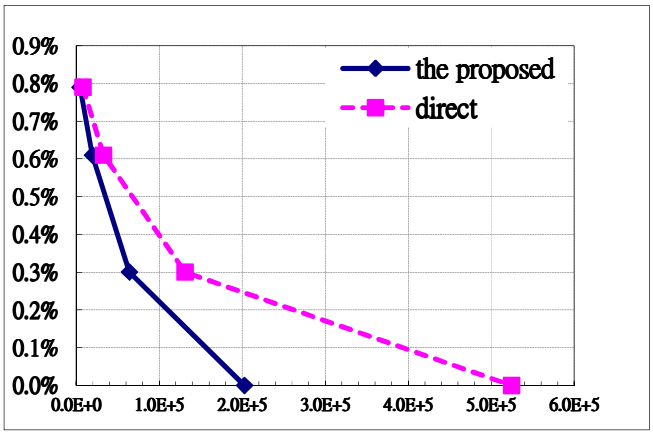


Fig. 6. The rate-distortion performance comparison in communication bandwidth for two methods.

The transmission bandwidth in our analysis is roughly computed by the data bits of the testing image because the bandwidth cost is proportional to the data bits transmitted in transmission. It is straightforward to count the data bits for the direct method, in which the picture is stored in the matrix form. In contrast, the data bits for the proposed method are computed from the node information in Table I. It is noted that the rate-distortion analysis for bandwidth can also be interpreted as the analysis for consumption rate for transmission power since the transmission power is usually proportional to the data rates. The processing power is approximated by the processing time, which is simply measured in average of 100 times and listed in Table II. As a summary, we can associate the processing time and data bits with the concepts of power cost and bandwidth cost respectively and have the corresponding two cost-distortion (rate-distortion) functions represented in Fig. 5 and 6.

V. CONCLUSIONS

In this paper, three strategies are proposed to improve the bit processing rate. First, utilizing the sensor's special feature, which emphasizes on decision not for visual aesthetics or entertainment purpose. Second, exploiting an appropriate functional image processing such as the morphological image processing, which provides the feature only (not the detail) for decision. Third, developing an image processing in the compressed form such as in the quadtree data structure, which is advantageous in the low resolution. Strategy 3 requires a technique innovation, image processing in the compressed form. It is noted that in the conventional approach the compressed image must be decompressed first to be processed. This innovation makes the compression and processing a joint design not two separate processes as before.

Morphological processing is suitable for fast module applications in sensor network because it is built upon simple logic operations with many extended functional

 ISBN: 978-1-61208-279-0

interpretations such as erosion, dilation, thinning, and pruning. The processing efficiency can be further improved by being implemented in hierarchical data structure. With its hierarchical data structure, the morphological processing can be applied to be focus on any specific blocks. The conventional direct method for morphological image processing is processed by a pixel vs. pixel basis with the pictures stored in the array form. In contrast, the morphological processing with a hierarchical data structure can be processed by a block vs. block basis. This joint processing concept is first time proposed in this paper. The processing efficiency is also verified in our experiments.

The rate-distortion theory is applied to our bandwidth gain analysis by defining the distortion as the error counts from the error images referenced to the target image computed by the best resolution. Meantime, the rate-distortion theory is further modified by us to study the power consumption rate if the rate is simply interpreted as the power consumption rate instead of the original bit rates. As a summary, it is the efficiencies in both the compression and the speed performance to make the proposed method outperform the direct method significantly in the rate-distortion analysis.

ACKNOWLEDGMENT

The authors would like to thank Chung Min-Fong and Gu Zheng-Fong for their helping the simulation and experiments in section IV. This research is partially supported by the National Science Council of the Republic of China under the contract NSC100-2221-E-110-080.

REFERENCES

[1] J. Sanchez, P. Ruiz, J. Liu, and I. Stojmenovic, “Bandwidth-efficient geographic multicast routing protocol for wireless sensor networks,” IEEE Sensors J., vol. 7, no. 5, pp. 627–636, May. 2007.

[2] R. Picard and T. Burr, “Networked sensors for cargo screening,” IEEE Sensors J., vol. 8, no. 8, pp. 1389–1396, Aug. 2008.

[3] H. Samet, Application of spatial data structure: computer graphics, image processing and GIS, New York: Addison-Wesley, 1990.

[4] J. Serra, Image Analysis and Mathematical Morphology, New York: Academic Press, 1983.

[5] J. W. Klingler, C. L. Vaughan, and L. T. Andrews, "Segmentation of echocardiographic images using mathematical morphology," IEEE Trans. on Biomedical Engineering, vol.35, No.11, pp. 925–933, 1988.

[6] J. M. Higgins, D. T. Eddington, S. N. Bhatia, and L. Mahadevan1, “Statistical dynamics of flowing red blood cells by morphological image processing,” PLoS Comput Biol., 5(2): e1000288. 2009.

[7] J. Velten and A. Kummert, “Implementation of a high-performance hardware architecture for binary morphological image processing operations;” Circuits and Systems, MWSCAS vol. 2, July, 2004, pp.241–244.

[8] J. C. Handley,”Minimal-Memory Bit-Vector Architecture for Computational Mathematical Morphology Using Subspace Projections,” IEEE Trans. on image processing, vol. 14, No. 8, Aug. pp. 1088–1096, 2005.

[9] C. Clienti, M. Bilodeau, and S. Beucher, “An efficient hardware architecture without line memories for morphological image processing,” Lecture Notes in Computer Science, vol. 5259, pp. 147–156, 2008.

[10] S. Chien, S. Ma, and L. Chen, “Partial-Result-Reuse architecture and its design technique for morphological operations with flat structuring elements,” IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 9, pp. 1156–1169, 2005.

[11] P. D. Gader, M. Khabou, and A. Koldobsky, "Morphological regularization neural networks," Pattern Recognition, Special Issue on Mathematical Morphology and Its Applications, vol. 33, no. 6, pp. 935–945, 2000.

[12] B. S. Chow,"A Limited Resources Based Approach to Coding for Wireless Video Sensor Networks,"IEEE Sensors Journal, Vol. 9, No. 9, pp. 1118–1124, Sep. 2009.

[13] P. Sussner and E. L. Esmi, “An introduction to morphological perceptrons with competitive learning,” in Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, June 14-19, 2009, pp. 3024–3031

[14] F. Meyer and J. Angulo, “ Micro-viscous morphological operators,“ in Proceedings of the 8th International Symposium on Mathematical Morphology, 2007, pp. 165–176.

[15] W. Wong, F. Y. Shih, and T. Su, “Thinning algorithms based on quad tree and octree representations,” Information Sciences, Vol. 176, Issue 10, 22, May, pp. 1379–1394, 2006.

[16] J. Voros, “Quad tree-based representations of grid-oriented data ,” Image and Vision Computing, Vol. 24, March, pp. 263–270, 2006.

[17] H. Shen, “Tracking method for embedded system,” Patent application IPC8 Class: AG06F944FI, USPC Class: 717128.

[18] V. Sklyarov, “ FPGA-based implementation of recursive algorithms,” Microprocessors and Microsystems. vol. 28, pp 197–211, 2004.

 ISBN: 978-1-61208-279-0

Driver Assistance System Towards Overtaking in Vehicular Ad Hoc Networks

Antonio S. S. Vieira, Joaquim Celestino Jr.
State University of Ceara (UECE)
Computer Networks and
Security Laboratory (LARCES)
Fortaleza, Ceara
Email: {sergiosviera, celestino}@larces.uece.br

Ahmed Patel, Mona Taghavi
Faculty of Information Science and Technology
University Kebangsaan Malaysia
Software Technology and
Management Research Center (SOFTEM)
Selangor, Malaysia
Email:{winchat2010, mona.taghavi}@gmail.com

***Abstract*—The large number of accidents caused by unsuccessful motor vehicle overtaking manoeuvres on roads is a significant problem with much public debate and concerns. In addition, one accident might cause another follow on accident when there is no signalling/reporting of it occurring, thus causing obstructions and further accidents in the vicinity. This can seriously affect the traffic flows, generating traffic jams and sometimes resulting in more unwarranted accidents, which can be avoided with the application of technology. The emergence of wireless ad hoc and sensor technologies and their adoption in vehicular networks can provide smart solutions to mitigate the probability of accidents and avoidance of dangerous traffic flow situations. This paper presents some important concepts for an application development to assist the driver in overtaking manoeuvres with prior knowledge of oncoming traffic, even when the road curvature has blind spots. In order to do that, the application must indicate whether it is safe or not to perform the overtaking manoeuvre. To propose this application, concepts of kinematics techniques are used to model the overtaking manoeuvres and the successful simulation results show this technique to be promising as expounded in this paper.**

***Keywords*-Driver Assistance System; Overtaking; VANET; Kinematics**

I. INTRODUCTION

Vehicular Ad Hoc Network (VANET) [1] is a new mobile communication system which promises many possibilities for a new range of applications in road traffic systems. Although VANETs are a newly introduced breed of innovative network concepts and technology application, many articles have been published about them, mainly from the point of view of safety applications [2], entertainment [3] and driver assistance applications [4].

These kind of applications can facilitate the human environment interaction in avoiding unwarranted mishaps/accidents. The driver assistance applications may be used, for example, to find parking spaces [5] and report on car accidents [6]. In the entertainment area, the human environment interaction application can quickly configure and harness players within a vicinity to join a multi-player on-line game [7].

In this paper, we present a driver assistance system for overtaking using VANET. We analyse the overtaking behaviour scenarios found in the literature [8] in order to capture our fundamental requirements to define our proposed maneuvering and overtaking model based on kinematic principles. This model decides through various algorithmic conditions when the overtaking manoeuvre is safe to be performed by a driver. This analysis is a very important cornerstone for designing a smart overtaking assistant, and it is the starting point of the first step for a complete and comprehensive VANET Driver Assistance System (VANET-DAS).

We also propose an intelligent message transmission technique in order not to overload the transmission medium since time is of the essence in VANET-DAS real-time activities with very tight threshold conditions and values. The decision-making for overtaking and maneuvering must be performed only in a real situation and the application must be able to identify a real overtaking situation as and when the events unfold. We also developed a rapid report message protocol which exchanges coordinates within the VANET vicinity for the surrounding vehicles to be context-aware of what is happening in a near instantaneous manner.

The remainder of this paper is organized as follows: Section II discusses the related work, Section III gives the theoretical analysis of the overviews the overtaking manoeuvre scenarios, Section IV describes the message broadcast system used by the applications within the VANET-DAS. In Section V, the experimental environment and the simulations results are presented and discussed; and finally, the conclusion and future works are given in the last section.

II. RELATED WORK

Nowadays, there are several proposed vehicular ad hoc network applications for safer driving and road traffic management, particularly concerning overtaking [9][10][11]. However, most of them related to overtaking do not consider the entire overtaking and maneuvering process. These applications mainly focuses on vehicular lane changing process, which in our work we regard as one of the many phases in an overtaking manoeuvre. Besides, not all of these related works do this in VANET environments.

Toledo et al. [9] developed a federated cooperative system to assist in overtaking on a cellular network. The overtaking is predicted after passing through filters that have applied kinematic information based on the vehicles and road topology format. The main purpose of Toledo's work is to estimate the risk in the overtaking manoeuvre. Similarly, our work also uses kinematics to predict a favourable overtaking manoeuvre situation, but with much more precision an explicitly in VANET environments and scenarios.

 ISBN: 978-1-61208-279-0

The main advantage of using VANET is that it can be formed either with or without physical infrastructure other than the air waves and frequency spectrum. Furthermore, in the application developed [9], vehicles exchange messages so that the manoeuvre can be authorized to be safe. During this operation many problems can arise such as sending replying messages transfer/transmission failure; lost message updates due to time-outs and vehicle traffic reconfiguration; loss of vehicle identification; and lack of updated management information about the VANET vicinity area in question. In this developed application, the messages are transmitted through broadcasting in a non-periodic manner and also when there is a change in the speed of the vehicle.

Ruder et al. [10] developed a system using wireless sensors in vehicles to assist the lane changing process in overtaking manoeuvres. The overtaking model in this system considers two lanes flowing in the same direction and an approximating vehicle system algorithm developed for predicting safe versus unsafe overtaking manoeuvres. Different from Rude's approach, we developed a coordinated positioning message broadcast protocol among the vehicles without the use of approximating wireless sensors. In our proposed VANET-DAS, there is the advantage of knowing the position of adjacent vehicles in advance even if they are a relatively long distance away since distance is defined by the transmission range.

Hrri et al. [11] developed an overtaking maneuvering application system which handles three different ways of area information recording and maintenance:

1) Constant Degree Detection: Every node tries to keep a constant number of neighbours. When a node detects that a neighbour actually left its neighbourhood, it tries to acquire new neighbours by sending a small advertising message.
2) Implicit Detection: A node i entering node $j's$ transmission range has a high probability to have a common neighbour with j.
3) Adaptive Coverage Detection: Every node sends an advertising message when it has moved a distance equal to a part of its transmission range.

In our proposed VANET-DAS, we developed a modified version of the Adaptive Coverage Detection technique. In Hrris scheme [11], it works by sending area messages reports at each $\frac{2}{3}$ (two thirds) [11] of the transmission range from the focal point. In our VANET-DAS version, the application checks in advance if an updated positioning message is necessary to be sent while comparing the real distance travelled and the predicted one. The calculations of the predicted distance is based on the last positioning message sent.

III. Theoretical Background and Analysis

According to Olsen [12], a lane change is defined as a deliberate and substantial shift in the lateral position of a vehicle, that is, when the vehicle leaves its original lane to manoeuvre towards another lane. According to Winsum et al. [13], a lane change happens in three sequential phases, considering the direction of the vehicle:

1) In the first phase, the car's steering wheel is turned to a maximum angle so that car can perform the the lane change.
2) The second phase starts when the wheel is turned in the opposite direction and ends when that angle reaches zero (always ahead).
3) During the third phase, the steering wheel is turned to a maximum angle in the opposite direction to establish the vehicle in the former lane.

Given the above definitions we can conclude that at least a minimum of two lane changes are involved in an overtaking manoeuvre. The first one shifts to one side to avoid colliding with the vehicle straight in front of it and the second one returns to the car's original lane.

According to Hegeman et al. [8] and Wilson et al. [14], an overtaking manoeuvre can be classified as:

1) Normal: The overtaker follows a vehicle and waits for a safe sufficient gap to perform an overtaking manoeuvre.
2) Flying: The overtaker does not adjust its speed to the speed of the vehicle that is to be overtaken but continues at its current speed during the overtaking manoeuvre.
3) Piggy backing: The overtaking vehicle follows another vehicle that overtakes a slower vehicle.
4) 2+: The overtaker performs the overtaking manoeuvre of two or more vehicles.

In our VANET-DAS, the overtaking scheme is based on the flying overtaking method. In the flying overtaking method (see Figure 1), the overtaking vehicle is travelling faster than the vehicle (leader) being overtaken. The faster vehicle changes lane, passes the leader and returns to its original lane without changing its speed except to ensure any distant vehicle coming from the opposite direction has sufficient distance to travel to avoid a collision (accident).

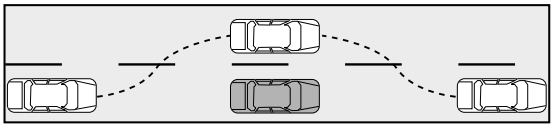


Figure 1. Overtaking using flying method

A. Kinematic-based Decision-Making

The overtaking decision takes into account the various combinations of collision possibilities with surrounding vehicles. That is, there is an appropriate and inappropriate scenario for performing the overtaking when the manoeuvre should or should not be performed according to a set of safety measures and criteria. Two characteristics of surrounding cars must be identified:

1) Position: The current position of the vehicle.
2) Mobility: The direction and speed of the vehicle.

The first characteristic, position, allows us to obtain a brief view on the network topology of the surrounding cars. In the

 ISBN: 978-1-61208-279-0

second characteristic, mobility, it is possible to go beyond the area knowledge of the topology, so that a network actor can predict a more precise topology configuration with more detailed information.

The task of mobility prediction is a challenge that has been solved through several techniques over time:

1) Adaptive Strategy: The predictions are corrected at the end of the average predictability interval.
2) Reactive Strategy: The node notifies the neighbourhood when a criterion changed (maximum predicting error).

Hrri et al. [11] details the evolution of such mobility prediction techniques and presents several schemes/models which are currently used nowadays in overtaking and maneuvering vehicle transport systems.

1) Deterministic Models: Only considers the position and a fixed velocity.
2) Stochastic Models: They do not aim at obtaining an exact prediction, but rather a correct one with high probability.

In our VANET-DAS, the trajectory of the car is obtained through a deterministic model, which is basically first order kinetics. Our application, having knowledge of a car, at the time (t), and in the present position (x,y) can show the vehicles future position using (1) [11].

$$Pos_i(t) = \begin{bmatrix} x_i + v_x^i.t \\ y_i + v_y^i.t \end{bmatrix} \quad (1)$$

This equation (1) calculates the distance between a vehicle and every other surrounding vehicle in the vicinity at an arbitrary time t, as given in (2):

$$\begin{aligned} D_{ij} &= D_{ji} \\ &= \|Pos_j(t) - Pos_i(t)\| \end{aligned} \quad (2)$$

The development of a high mathematical fidelity model can be very complex, depending on the various variables/coefficients involved, such as speed, acceleration, air resistance, weather condition, road condition, other drivers behaviour patters and the driver's reaction time. In a real, practical, overtaking manoeuvre scenario, several factors must be considered. In our work we focus on a few essential variables and conditions namely, position and mobility, resulting in a simplified model but with equally valid outcomes.

The scenario in which the overtaking model was developed involved three vehicles, C_1, C_2 and C_3 as shown in Figure 2.

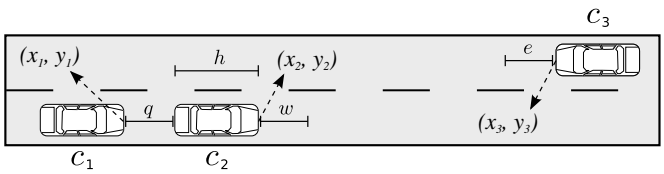


Figure 2. Scenario used in the simplified overtaking model

C_1 is travelling faster than the speed of C_2 and C_3, and their individual speeds are non-zero values. Looking at Figure 2, q, w and e are constants, and given the relative speeds, they refer to the safe distances of the respective vehicles, while h refers to the average length of an arbitrary vehicle. The current position of each vehicle is typically represented by (x_i, y_i) as shown in Figure 2 and their velocity is represented by (v_{xi}, v_{yi}), with the index i identifying each vehicle.

For overtaking decision making, it is necessary to calculate the travelled distance of the whole overtaking manoeuvre. At the end of the manoeuvre, the current car position is validated for its position and verified to identify if the manoeuvre can take place successfully within the time constraints or threshold value bounds.

As previously mentioned, in the flying model, the vehicle does not accelerate for overtaking, although in practice this would be deemed necessary if either the vehicle being overtaken decides to accelerate or the oncoming vehicle is travelling faster than expected/predicted. In our scenario, the overtaking car has a higher maximum speed than the leader and performs the overtaking manoeuvre with a constant speed. The manoeuvre itself is performed in two stages.

First Stage: there are four steps in this stage. The first step is when the C_1 vehicle is behind the leader and its first manoeuvre is positioning itself to perform lane shifting as they travel in tandem. In other words, it must travel l distance on y-axis in a somewhat diagonal lane shift manoeuvre as shown in Figure 3.

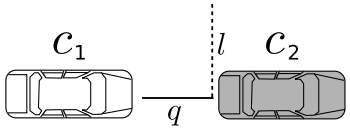


Figure 3. Distance travelled on lane shift

The second step is to move up for overtaking, C_1 needs to rotate the vehicle steering wheel making a relative angle θ between the wheels and the x-axis as shown in Figure 4. The vehicle drifts to shift into the adjacent lane. Knowing this information, the velocity components v_x^{c1} and v_y^{c1} are calculated using equations 3 and 4 respectively:

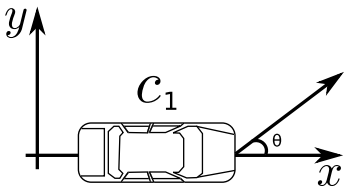


Figure 4. Angle between the wheels and x-axis on lane shift

$$v_x^{c1} = v_{c1} \cdot cos(\theta) \quad (3)$$

$$v_y^{c1} = v_{c1} \cdot sin(\theta) \quad (4)$$

The 3rd step is to calculate the time (t_1) and distance travelled (s_1) by C_1 during the lane shift. This is calculated using 5 and 6 respectively:

$$t_1 = \frac{l}{v_y^{c1}} \tag{5}$$

$$s_1 = v_x^{c1} \cdot t_1 \tag{6}$$

In the final step, it is necessary to calculate the remaining C_1 vehicle's distance (r) so that the C_1 vehicle's front is paired with the C_2 vehicle's back as shown in Figure 5. This is calculated using 5.

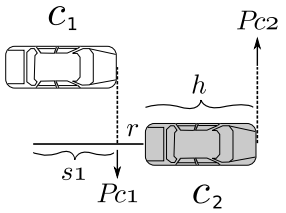


Figure 5. Remaining distance of C_1 so that C_1's front is paired with the back of C_2

$$r = \|Pos_{c2}(t_1) - Pos_{c1}(t_1)\| - h \tag{7}$$

Second Stage: C_1 must pass C_2. In order to achieve this, C_1 has to travel a minimum of $2 \cdot r + 2 \cdot h$ distance as shown in Figure 6.

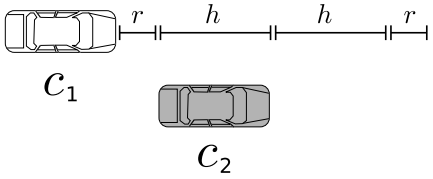


Figure 6. Distance travelled by C_1 to safely pass C_2

The time spent in the second stage overtaking manoeuvre (t_3) is calculated using (8):

$$t_3 = \frac{2 \cdot r + 2 \cdot h}{\|\vec{v}_{c1} - \vec{v}_{c2}\|} \tag{8}$$

and the travelled distance (s_3) is calculated using 9:

$$s_3 = \vec{v_{c1}} \times t_3 \tag{9}$$

For C_1 to return to its original lane side, the same values are used (time and travelled distance) that are calculated during the first stage overtaking manoeuvre move (Figure 7).

After obtaining s_1 and s_3, the total travelled distance by C_1 during the overtaking manoeuvre is calculated by using $s_t = 2 \cdot s_1 + s_3$.

To check if the overtaking can be performed safely, the time must be calculated so that C_1 will be in a safe position with

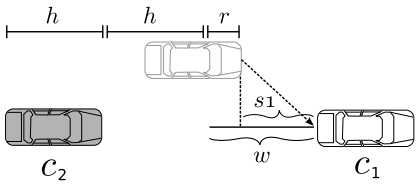


Figure 7. Travelled distance for C_1 returns to its original lane

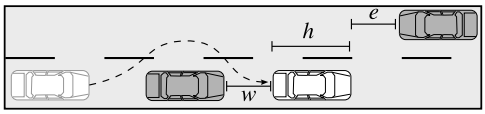


Figure 8. Desired position of C_1 after it complete the whole overtaking manoeuvre

sufficient leeway to move into the original side of its lane with a drift of θ degrees as shown in Figure 8.

To achieve this drifting position in a safe manner with collision avoidance, the following equation (10) must be solved:

$$t_4 = \frac{\|Pos_{c2}(0) - Pos_{c3}(0)\| - w - h - e}{\|\vec{v_{c2}} + \vec{v_{c3}}\|} \tag{10}$$

such that the distance travelled by vehicle C_1 is calculated (s_4) during the overtaking manoeuvre in t_4 seconds.

For making the kinematic-decision for safe collision-free overtaking manoeuvre, s_t and s_4 are compared. If the s_t value is lower than the s_4 value, overtaking is allowed to proceed safely within the threshold bounded values of all the other related variables/parameters.

B. Overtaking Situation Detection

The detection of the overtaking situation is important in our VANET-DAS. It is a trigger for the calculating process to validly detect the overtaking situation using the kinematic decision-making principle based on three conditions:

1) Overtaking intention.
2) Passing vehicle is really behind the leader at a safe distance.
3) Check if the vehicles are travelling in the same direction.

The first condition refers to the overtaking intention value represented by γ (gamma). For each vehicle in our VANET-DAS, there is an associated value for the overtaking intention over a period of time. The γ value is represented by a real number in the interval $[0, 1]$, where 0 denotes null overtaking intention and 1 denotes maximum overtaking intention.

A vehicle overtaking intention γ_i at time t is calculated using (11).

$$\gamma_i(t) = \frac{E}{D_{ij}(t) + E} \tag{11}$$

where $D_{ij}(t)$ refers to the distance of vehicle i to the closest vehicle j and E refers to the sum of the safety distances q and the length h of a vehicle, in other words, $E = q + h$.

 ISBN: 978-1-61208-279-0

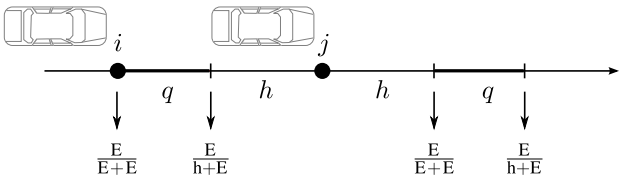


Figure 9. Range of values for overtaking intention

This condition is true when $\frac{E}{E+E} \leq \gamma \leq \frac{E}{h+E}$ holds, and it is equal to the line in bold represented in 9.

When the vehicle i is at a safety distance of E meters from vehicle j, the overtaking intention will be $\frac{E}{E+E} = \frac{E}{2E} = 0.5$ and when i is at h meters of j the overtaking intention will be the maximum safety limit allowable. For example, when when $h = 8.0$ and $q = 33.3$, $\gamma = \frac{33.3+8.0}{8.0+33.3+8.0} = 0.83$.

The second condition is used to ensure that the overtaking is really behind the leader at a relatively safe distance. For this condition, it is necessary to calculate the angle between the vehicles as shown in Figure 10.

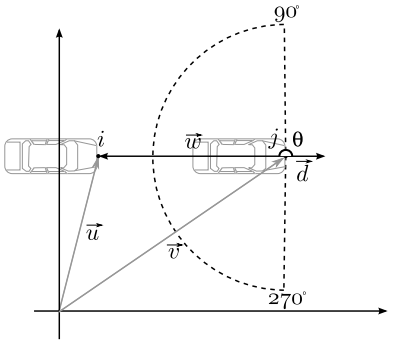


Figure 10. Angle between direction vector of j and vector w

In Figure 10, $\vec{u}$ and $\vec{v}$ are positioning vectors of vehicles i and j respectively and $\vec{w}$ (vector difference) is derived from $\vec{w} = \vec{u} - \vec{v}$. Therefore, to ensure that vehicle i is behind vehicle j, the angle between $\vec{w}$ and $\vec{d}$ should be in the 90 to 270 degree range, i.e., $90 \leq \theta \leq 270$.

The third (last final) condition is used to check if the vehicle i and j are travelling in the same direction. For this condition, it is sufficient to check the angle between the direction vectors of i and j as illustrated in 11.

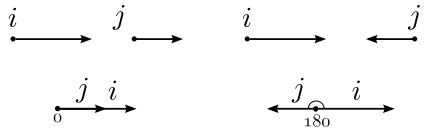


Figure 11. Angle between vector direction of vehicle i and j

If the angle between direction vectors of i and j is between the interval $[\theta, 360 - \theta]$, as shown in Figure 12, the last condition is satisfied.

Finally, it can be stated that if all three of the above mentioned conditions are fulfilled, then a valid safe overtaking maneuvering situation holds true.

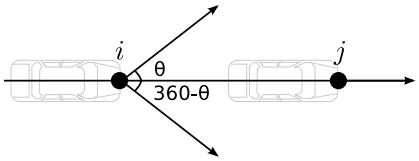


Figure 12. Interval for identifying when two vehicles are travelling in same direction.

C. Positioning Messages

In a vehicular safety application in a VANET environment, the message exchange among vehicles is an important operation, which itself must be very explicit to avoid any misrepresentations and misinterpretations. Each device within the vehicle must be able to predict a safe situation for an overtaking move. For this to happen without overloading the network with unnecessary network transmission traffic, it is necessary to ensure that the transmission medium is used in a smart and efficient manner. In other words, the application should use the transmission medium only when necessary within very stringent/tight command and response round-trip delay threshold value bounds. For this, we are using a scheme in which the predicted position is compared periodically to the real-life position. If there is a significant difference between these values, a new message is broadcasted to all surrounding vehicles with in its vicinity range.

When a vehicle is moving, the distance to be covered before broadcasting a mobility message (since vehicles are typically dynamically moving) is defined how $\frac{2}{3}$ (two thirds) of the transmission range. With this distance and the vehicle speed, one can easily predict the time during which a new broadcast should be performed (that is predicting t). The interval between the initial time and the predicted time is divided into seven equal time intervals as shown in Figure 13.

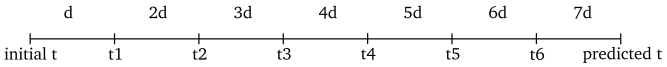


Figure 13. Intervals of time for checking if it is necessary to send a new positioning message.

After each interval t_i, a comparison between each vehicle's real position (P_a) and its predicted position (P_p) is performed, and if the absolute difference between P_a and P_p is greater than a sufficiently small value, ϵ (epsilon, ϵ = 0.5), a new message is broadcasted to all the vehicles in the vicinity. This scheme of the broadcasting method has been devised to ensure that our VANET-DAS does not broadcast unnecessary messages by consuming transmission and device processing capacities. In overall system and network management terms remain efficient within the constraints of the dynamic real-time environment of a VANET.

IV. Simulations and Results

The experiments were performed using Network Simulator ns2, version 2.34 [15] using an Intel Core i3 CPU 2.13 GHz

 ISBN: 978-1-61208-279-0

computer with GNU-Linux Ubuntu 11.04 [16]. Our VANET-DAS application was modelled and developed as a set of mobile agents to reflect the vehicles in a transport environment as shown in Figure 14. Each device in then simulation used Nakagami propagation model [17] and $1,000$m transmission range.

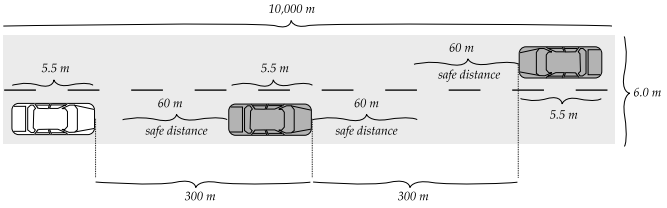


Figure 14. Simulated real-life like environment

The surrounding vehicles data is stored in an optimized table-lookup data structures, which provide both rapid searchs and update using C++ programming with the Standard Template Library (STL) [18] functions for it.

The mobile agents representing the vehicles (devices of a VANET) were established through a common higher-level controlling meta-agent which generates a log file that provides data about safe and unsafe overtaking situations for each of the vehicles in the experiment. This log file is filtered to identify the devices in each vehicle during the overtaking maneuvering and inter message exchange operations being simulated.

We analyse and evaluate the data collected from the simulation runs to assess the vehicles' behaviour for two functionalities of our VANET-DAS:

1) Broadcasting
2) Overtaking Situation Detection

And finally, we also analyse the reliability of the VANET-DAS application itself. The results are given below.

A. Broadcasting Results

As shown in Figure 15 it is possible to identify that the VANET-DAS application works well, for example, when a vehicle moves with a constant 20 m/s speed during 100 s and $1,000$ m transmission range, the application sends a report message at each 666.67 m.

In this case, three messages are sent when the vehicle is at 666.67 m, $1,333.33$ m and $2,000$ m respectively, stating from an initial position and a single last message is sent in the initial simulation point, totalling four messages. Similarly, behaviour is also identified in our other simulation runs. Therefore, the VANET-DAS application has an expected behaviour because it sent positioning messages at expected positions.

B. Overtaking Situation Detection Results

In this experiment, we observed and checked the overtaking detection operation using our VANET-DAS simulated scenario where a vehicle (C_1) is 155 m from another vehicle (C_2) and they are 32 m/s and 14 m/s apart respectively. C_1 is approaching C_2 position and is approximately 41.3m behind when C_1 starts the overtaking manoeuvre in three stages:

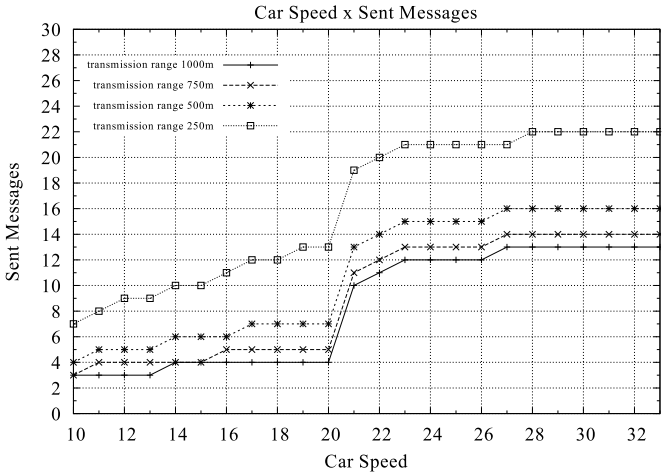


Figure 15. Amount of sent messages as a function of vehicle's speed and transmission range

1) Lane shift;
2) Overtaking; and
3) Returning to the original lane

Figure 16 shows the detection of time of the overtaking manoeuvres. At x-axis, its pointed is determined as 0 when the method does not detect an overtaking, and it will be 1 when an overtaking is detected.

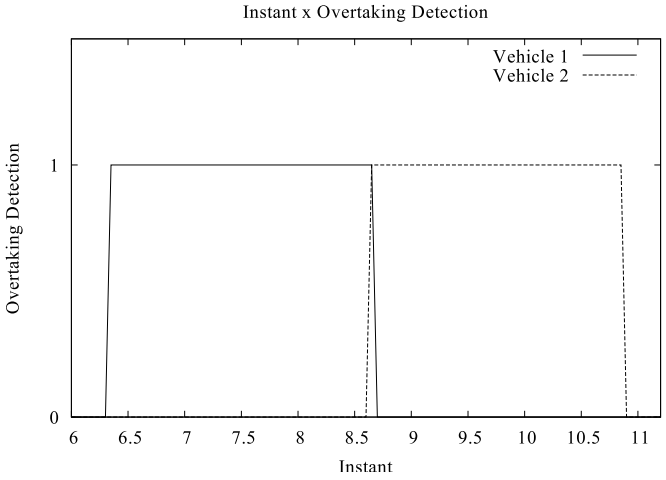


Figure 16. Instants which has detected overtaking situation

In order to demonstrate and validate if the results obtained in the experiment are consistent, as illustrated in Figure 16, we used 14 to calculate the instant (horizontal axis) when the vehicle C_1 is q meters close to vehicle C_2:

$$S_{C1}(T) = S_{C2}(T) - q - h \tag{12}$$

$$x_1 + vx_1T = x_2 + vx_2T - q - h \tag{13}$$

$$T = \frac{x2 - x1 - q - h}{vx_1 - vx_2} \tag{14}$$

Defining $q = 33.3$ and $h = 8.0$, we get $T = 6.31$ seconds, we notice in Figure 16 that at this point an overtaking situation

was detected.

C. Reliability of the Application

The main goal in this experiment is to validate the application operation in different scenarios of our VANET-DAS. In each scenario, the vehicle C_1 is placed randomly behind C_2 and C_3 placed randomly in front of C_2 as shown in Figure 17.

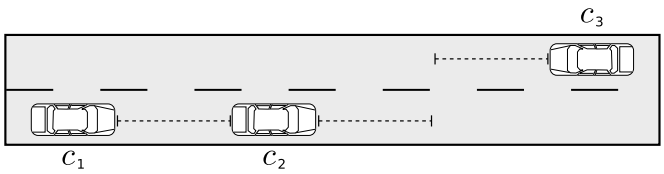


Figure 17. A random placement of vehicles simulation scenario

The vehicle's speed is chosen randomly as follow:

1) C_2's speed is chosen between the range of 16 m/s and 25 m/s.
2) C_1's speed is chosen between the range of C_2s speed plus 5 m/s and 30 m/s.
3) C_3's speed is chosen between the range of 16 m/s and 30 m/s range.

Using these parameters, $1,000$ different scenarios were created and in each scenario the time in which C_1 passes C_3 (t_c) is checked. To confirm the application accuracy, two times are identified in the overtaking manoeuvre. The first time detection happens when C_1 is behind C_2 at a distance of $q+h$ meters (t_1) and the second time detection when C_1 is in front of C_2 at the distance of $h+q$ meters (t_2) as shown in Figure 18.

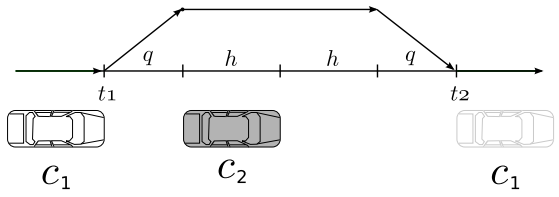


Figure 18. The times t_1 and t_2 times in overtaking manoeuvre

Thus, we can safely conclude that the VANET-DAS application should work correctly when any one of the following conditions are met:

1) If t_c is lower than t_1 it is expected that the application will not send any message because it has not detected an overtaking situation.
2) If t_c is between t_1 and t_2 the application is expected to indicate that an overtaking cannot be performed safely.
3) If t_c is greater than t_2 one can expect that the application indicates that overtaking can be safely performed.

At each iteration of the experiment, we compared the instant t_c with times which the application generates alerts of overtaking permission or prohibition. We found that the application returned the expected messages in 99% of the cases in our experiments.

V. Conclusion and Future Works

The research work discussed in this paper presented some fundamental concepts for an application development in the VANET environment to assist drivers in overtaking manoeuvres safely. Using kinematics and a positioning mechanism with an efficient messaging communication protocol, it is possible to develop a real-life application system capable of predicting safe overtaking manoeuvres with minimum near collision or no accident risk and without overloading the VANET data transmission medium.

Despite using controlled scenarios and without considering the driver's unpredictable behaviour pattern, the application has promising characteristics for an actual application development to assist overtaking in real-life systems for modern traffic systems. In particular, as shown from the result, the application can detect an ongoing overtaking manoeuvre operating in an efficient message data transmission VANET environment with a near 99% reliability.

As shown from the result, the application can detect an ongoing overtaking manoeuvre saving CPU processing. Moreover, the system could predict 99% of simulated situations.

In our future research work, we intend to improve the broadcast mechanism analysing the balancing act between position checking interval and the positioning errors in which the checking intervals are not constant but dynamically set according to the vehicle positioning and vehicle space prediction error margins and travelled distance between message broadcast periods. It is considered a difficult problem to solve because it is desirable to increase the travelled distance by the vehicle before sending a new message report to save on transmission overheads. This requires a critical balancing act between the two conflicting goals of safe distance between vehicles and active communication (without loss of message communication) in the VANET vicinity transmission range. It is also necessary to reduce the positioning error between the predicted and real vehicle positions.

References

[1] R. Morris, J. Jannotti, F. Kaashoek, J. Li, and D. Decouto, "CarNet: A scalable ad hoc wireless network system," in *Proceedings of the 9th workshop on ACM SIGOPS European workshop: beyond the PC: new challenges for the operating system*. ACM, 2000, pp. 65.

[2] W. Chen and S. Cai, "Ad hoc peer-to-peer network architecture for vehicle safety communications," *Communications Magazine, IEEE*, vol. 43, no. 4, pp. 100-107, 2005.

[3] H. Yoon, J. Kim, F. Tan, and R. Hsieh, "On-demand video streaming in mobile opportunistic networks," in *Proceedings of the 2008 Sixth Annual IEEE International Conference on Pervasive Computing and Communications*. IEEE Computer Society, 2008, pp. 80-89.

[4] L. Wischoff, A. Ebner, H. Rohling, M. Lott, and R. Halfmann, "SOTIS-a self-organizing traffic information system," in *Vehicular Technology Conference, 2003. VTC 2003-Spring. The 57th IEEE Semiannual*, vol. 4, 2003.

[5] M. Caliskan, D. Graupner, and M. Mauve, "Decentralized discovery of free parking places," in *Proceedings of the 3rd international workshop on Vehicular ad hoc networks*. ACM, 2006, pp. 39.

[6] S. Rahman and U. Hengartner, "Secure crash reporting in vehicular ad hoc networks," in *Proceedings of the 3rd International Conference on Security and Privacy in Communication Networks (SecureComm2007)*. Citeseer.

 ISBN: 978-1-61208-279-0

[7] O. Tonguz and M. Boban, "Multiplayer games over VANET: a new application," *Ad Hoc Networks*, 2010.
[8] G. Hegeman, K. Brookhuis, S. Hoogendoorn *et al.*, "Opportunities of advanced driver assistance systems towards overtaking," *EJTIR*, vol. 5, no. 4, pp. 281-296, 2005.
[9] R. Toledo-Moreo, J. Santa, and M. Zamora-Izquierdo, "A cooperative overtaking assistance system," *Planning, Perception and Navigation for Intelligent Vehicles (PPNIV)*, pp. 50-56, 2009.
[10] M. Ruder, W. Enkelmann, and R. Garnitz, "Highway lane change assistant," in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 1. IEEE, pp. 240-244.
[11] J. Härri, "Modeling and predicting mobility in wireless ad hoc networks," Ph.D. dissertation, lÉcole Polytechnique Fédérale de Lausanne (EPFL), 2007.
[12] E. Olsen, "Modeling slow lead vehicle lane changing," Ph.D. dissertation, Citeseer, 2003.
[13] W. Van Winsum, D. De Waard, and K. Brookhuis, "Lane change manoeuvres and safety margins," *Transportation Research Part F: Psychology and Behaviour*, vol. 2, no. 3, pp. 139-149, 1999.
[14] T. Wilson and W. Best, "Driving strategies in overtaking." *ACCID. ANALY. & PREV.*, vol. 14, no. 3, pp. 179-185, 1982.
[15] S. McCanne, S. Floyd, K. Fall, K. Varadhan *et al.*, "Network simulator ns-2," 2000.
[16] M. Shuttleworth, "Ubuntu: Linux for human beings," 2011.
[17] M. Nakagami, K. Tanaka, and M. Kanehisa, "The m-Distribution As the General Formula of Intensity Distribution of Rapid Fading," *Memoirs of the Faculty of Engineering, Kobe University*, vol. 4, pp. 78-125, 1957.
[18] D. Musser and A. Saini, *The STL Tutorial and Reference Guide: C++ Programming with the Standard Template Library*. Addison Wesley Longman Publishing Co., Inc. Redwood City, CA, USA, 1995.

 ISBN: 978-1-61208-279-0

Indoor Localization for Multi-Wall, Multi-Floor Environments in Wireless Sensor Networks

Xiao Fan and Yoan Shin
School of Electronic Engineering
Soongsil University
Seoul, Korea
{xiaofan; yashin}@ssu.ac.kr

***Abstract*—In recent years, importance of indoor localization has increased in Wireless Sensor Networks (WSNs). Extremely complex nature of the indoor environments makes WSN research more difficult than in the outdoor counterparts. In WSNs, the Received Signal Strength (RSS) from sensors is usually utilized to locate unknown sensors or nodes. However, the RSS is relatively sensitive to indoor obstructions such as walls and ceilings between adjacent floors. Various indoor localization algorithms are available, but most are not suitable because of the complex indoor environments. This paper proposes a novel localization scheme which uses vertical allocation of beacon sensors to estimate floor attenuation and wall attenuation. These two parameters are then used to identify the floor number and to estimate the horizontal localization of an unknown node, respectively. Simulation results show that both schemes provide accurate localization performance.**

Keywords—Wireless sensor network; Indoor localization; Multi-floor; Multi-wall; Attenuation; Received signal strength

I. INTRODUCTION

With the development of wireless networks and digital communication devices, people are paying increasing attention to location-based services. These can be applied in many fields such as vehicle systems, tourist guides, and personal security, to name a few. The Global Positioning System (GPS) is a mature and widely used technique, especially in outdoor positioning or localization, due to its high accuracy [1]. Unfortunately, GPS is not suitable for large-scale sensor networks because of its high hardware costs and energy requirements. More importantly, GPS is not suitable for use indoors because of its low penetrating satellite signals. Hence, indoor localization technology is more challenging, and it has become an important focus of research interest in recent years [2].

The availability of Wireless Sensor Networks (WSNs) has contributed to indoor localization [3]. In general, a WSN is composed of numerous low-power sensors or nodes, which are designed to be a small computing unit. A WSN is usually equipped with different kinds of sensors to monitor targets and communicate with each other. Due to its small size and low-cost nature, a large number of sensors can be easily used to locate the target in a building. The localization method proposed in this paper is suitable for indoor environments with WSN with multiple floors and multiple walls.

The remainder of this paper is organized as follows. Section II discusses previous work in localization for WSN. Section III describes the system model including the estimation of the attenuation of the floors and the walls. In Section IV, we propose a localization scheme which utilizes the estimated attenuation values from the previous section to identify the floor number and the localization of the plane. Section V presents an improved approach to increase the localization accuracy. The simulation results are given in Section VI, followed by the concluding remarks in Section VII.

II. RELATED WORKS

A variety of localization methods for WSNs have been proposed. In relation to the locations of nodes, those for which the coordinates are known in advance are called beacon nodes. If the location is not known in advance, they are called unknown or target nodes. Existing localization algorithms are largely divided into two typical categories depending on the distance between beacon node and unknown node: range-based [4] and range-free [5]. The former is defined by protocols that use absolute point-to-point distance (or range) estimates or angle estimates for estimating the location. The latter uses network connectivity and other information to localize the nodes. There are some typical algorithms such as the centroid algorithm, Approximate Point in Triangle Test (APIT), DV-hop and so on. However, a large number of nodes are required for effective range-free localization.

In range-based schemes, Received Signal Strength (RSS) based localization is less expensive and simple to implement in hardware. Thus, it is widely used as compared to other methods such as Time of Arrival (ToA), Time Difference of Arrival (TDoA), and Angle of Arrival (AoA) schemes. However, the RSS scheme is very sensitive to fading effects and multipath reflection. Numerous methods are proposed to solve this problem, although their effects are not satisfactory. Generally, a

 ISBN: 978-1-61208-279-0

proper propagation model suitable for complex indoor environment is significant and some models are also proposed. The Multi-Wall (MW) model proposed in [6] only considers attenuation caused by wall number and its texture. However, the number of walls should be known in advance and the floor attenuation in multi floor building is not also considered. These indoor obstructions such as walls, material layers will result in low localization accuracy. To address the issue caused by obstruction, this paper designs a sensor-based localization scheme in which the sensors are aligned vertically. This scheme takes the floor and wall attenuation into consideration. Also, it is mainly analyzed by using a simple application example.

III. SYSTEM MODEL

A. Allocation of Beacon Nodes

With respect to the allocation of the beacon nodes, each floor is supposed to be equipped with the same number of nodes which have the same horizontal coordinates. Thus, each beacon node and its adjacent beacon node upstairs and downstairs comprise a vertical group. This group is used to estimate the floor attenuation and the wall attenuation in the following scheme. Fig.1 shows this type of allocation of the beacon nodes. In this paper, four vertical groups are set at the corner of the plane area and H is the distance between adjacent beacon nodes in the same vertical group.

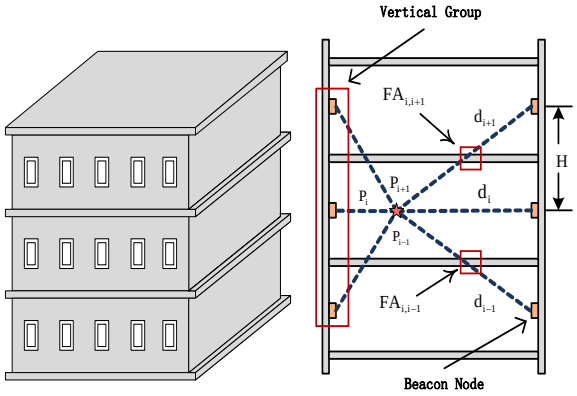


Figure 1. Allocation of vertical beacon nodes

B. Floor Attenuation Estimation

The floor attenuation is estimated as shown in Fig. 1 where $FA_{i,i+1}$ denotes the floor attenuation between the i-th floor and the i+1 -th floor. Based on a conventional RSS model [7] as given in (1), a new RSS model for floor attenuation estimation is proposed by (2).

$$PL(d) = PL(d_0) + 10\mu \log_{10}(d / d_0) + X_s \quad (1)$$

$$PL(d) = PL(d_0) + 10\mu \log_{10}(d / d_0) + FA + X_s \quad (2)$$

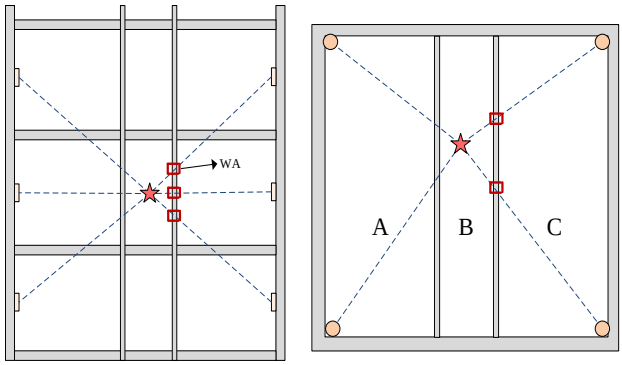
where d is the distance between the transmitter and the receiver, d_0 is the reference distance, μ is the path loss exponent, FA is the floor attenuation, and X_s is a zero-mean normally distributed random variable.

C. Wall Attenuation Estimation

Based on the floor attenuation model, we consider a wall attenuation during propagation as shown in (3).

$$PL(d) = PL(d_0) + 10\mu \log_{10}(d / d_0) + FA + nWA + X_s \quad (3)$$

where n is the number of walls during the propagation and WA is the single wall attenuation. Note that WA needs to be measured in advance. Here, as a simple illustrative case shown in Fig.2 (b), two vertical walls that divide the horizontal plane into three areas are set in the building.



(a) (b)

Figure 2. Wall attenuation estimation model and plane positioning

Actually, this wall attenuation model proposed here is suitable for the buildings with the same room layout in each layer. That will guarantee that the signals will suffer from the same wall attenuation from the same vertical group between adjacent layers.

IV. PROPOSED LOCALIZATION SCHEME

A. Floor Number Identification

For a vertical group, we can transform the measured RSS attenuations from the middle beacon node PL_i, the upstairs beacon node PL_{i+1}, or the downstairs beacon node PL_{i-1} to the distance using (1) and (2), respectively. Then, (4) and (5) can be obtained. According to the geometry theory, the following relationship (6) can be also derived.

$$d_i = 10^{\frac{PL_i - PL(d_0)}{10u}} \quad (4)$$

$$d_{i+1/i-1} = 10^{\frac{PL_{i+1}/PL_{i-1} - FA_{i+1}/FA_{i-1} - PL(d_0)}{10u}} \quad (5)$$

$$H^2 + d_i^2 = d_{i+1}^2 = d_{i-1}^2 \quad (6)$$

Then, by inserting (4) and (5) into (6), the floor attenuation between the i-th floor and the i+1 -th floor

$FA_{i,i+1}$ or $FA_{i,i-1}$ between i-th floor and i-1-th floor can be calculated by (7).

$$FA_{i,i+1/i-1} = P_{i+1/i-1} - PL(d_0) - 5u\log_{10}(H^2 + 10^{\frac{PL_i - PL(d_0)}{5\mu}}) \quad (7)$$

If the floor attenuation for each floor is obtained through *a priori* experimental measurement, the estimated floor attenuation with (7) can be employed, together with Minimum Euclidean Distance Matching (MEDM) as shown in (8) to identify the location of the floor of the unknown node. $PFA_{i,i+1}$ is the priori floor attenuation between the i-th floor and i+1-th floor. For instance, the estimated i-th floor attenuation value is shown in the last column of Table 1. After the matching process with the second floor attenuation and the fourth floor attenuation, respectively, the estimated floor of the unknown node is equal to 4, depending on the minimal matching value shown in Table 1.

$$MEDM = \sqrt{(FA_{i,i-1} - PFA_{i,i-1})^2 + (FA_{i,i+1} - PFA_{i,i+1})^2} \quad (8)$$

TABLE 1. FLOOR NUMBER IDENTIFICATION USING ESTIMATED ATTENUATION VALUES

Estimated value	Floor number		
	2	4	i
$\boldsymbol{FA_{i,i-1}}$	$PFA_{2,1} = 15.4\,dB$	$PFA_{4,3} = 18.7dB$	$FA_{i,i-1} = 17.9dB$
$\boldsymbol{FA_{i,i+1}}$	$PFA_{2,3} = 18.7dB$	$PFA_{4,5} = 16.9dB$	$FA_{i,i+1} = 17.3dB$
MEDM	2.87	0.89	$i = 4$

This floor identification approach can achieve a high accuracy, especially where there is a large difference in the floor attenuation between the different floors. In practical terms, it can be applied in buildings with significant floor structure differences.

B. Plane Positioning

Due to similar structural composition and distribution of the walls in each floor of the building, the signal is likely to suffer from the same wall attenuation from the same vertical group between two adjacent floors. Let WA_j be the wall attenuation during the propagation from the j-th vertical group. According to the geometry theory used in floor number identification, we have

$$d_i = 10^{\frac{PL_i - PL(d_0) - WA_j}{10u}} \quad (9)$$

$$d_{i+1/i-1} = 10^{\frac{PL_{i+1}/PL_{i-1} - FA_{i+1}/FA_{i-1} - WA_j - PL(d_0)}{10u}} \quad (10)$$

$$WA_j = -5u\log_{10}\left(\frac{H^2}{10^{\left(\frac{PL_{i+1/i-1,\ j} - FA_{i,i+1/i-1} + PL(d_0)}{5\mu}\right)} - 10^{\left(\frac{PL_{i,j} - PL(d_0)}{5\mu}\right)}}\right) \quad (11)$$

where $PL_{i,j}$ is the measured RSS attenuation from the i-th floor's beacon node of the j-th vertical group.

After the wall attenuation from each beacon node has been estimated using (11), the location of the unknown node can be evaluated. However, *a priori* information on the distribution of the walls should be obtained in advance as shown in Fig. 2(b). For convenience, a wall number matrix, $WN^{N\times M}$ is defined as shown in (12).

$$\begin{bmatrix} 0 & 2 & 0 & 2 \\ 1 & 1 & 1 & 1 \\ 2 & 0 & 2 & 0 \end{bmatrix} \quad (12)$$

where N is the number of areas and M is the number of beacon nodes for each layer. Actually, M is also the number of vertical groups in every three layers.

The estimated wall attenuation can be transformed into the estimated wall number by utilizing the measured WA in (3). Based on that, we can know whether there exists walls and how many walls along each propagation path. Then, we can match the estimated wall number with each row of the prior wall number matrix $WN^{N\times M}$ one by one using the MEDM. In that case, the area of the unknown node will be accurately determined.

To achieve a high localization accuracy, the wall attenuation of each floor where the unknown node is located should be eliminated. Here, we use two methods to revise the measured RSS of the unknown node. One involves direct utilization of the estimated wall attenuation, and the other entails using WA and the estimated wall number.

After we revise the RSS, the conventional plane algorithm can be applied to the plane localization. For the triangle method given in [8], the RSS of at least three beacon nodes should be selected. Using the conventional RSS model, we can transform the RSS into the distance and put it into the plane localization. Here, we denote (x, y), (x_1, y_1), (x_2, y_2), (x_3, y_3) as the locations of the unknown node and the three selected beacon nodes. Then, for the selected three beacon nodes on the horizontal planes, we denote $d_{i,1}$, $d_{i,2}$, and $d_{i,3}$ as the distances between the unknown and the three selected beacon nodes on the i-th floor. Using (13) and (14), the estimated coordinates of the unknown node can be obtained as

$$\begin{aligned} \alpha &= (d_{i,1}{}^2 - d_{i,2}{}^2) - (x_1{}^2 - x_2{}^2) - (y_1{}^2 - y_2{}^2) \\ \beta &= (d_{i,1}{}^2 - d_{i,3}{}^2) - (x_1{}^2 - x_3{}^2) - (y_1{}^2 - y_3{}^2) \end{aligned} \quad (13)$$

$$x = \frac{\begin{vmatrix} \alpha & 2Y_1^2 \\ \beta & 2Y_1^3 \end{vmatrix}}{\begin{vmatrix} 2X_1^2 & 2Y_1^2 \\ 2X_1^3 & 2Y_1^3 \end{vmatrix}} \quad y = \frac{\begin{vmatrix} 2X_1^2 & \alpha \\ 2X_1^3 & \beta \end{vmatrix}}{\begin{vmatrix} 2X_1^2 & 2Y_1^2 \\ 2X_1^3 & 2Y_1^3 \end{vmatrix}} \quad (14)$$

where X_b^a and Y_b^a refer to $(x_a - x_b)$ and $(y_a - y_b)$.

 ISBN: 978-1-61208-279-0

In fact, the location of the unknown node can be obtained with $\binom{M}{3}$ selections of the RSSs. However, preference should be given to an RSS that experiences less wall attenuation or even no attenuation. Although we have revised the RSS by eliminating the wall attenuation, error always exists between the real wall attenuation and the estimated wall attenuation. Here, we can select the beacons depending on the smallest three wall number values for the corresponding row of $WN^{N\times M}$. In a best-case scenario, there are three RSSs without wall attenuation. That is to say, there should be at least three zeros for some rows of $WN^{N\times M}$ in that case.

V. IMPROVED LOCALIZATION METHOD

The method given in Section III uses three distances from three beacon nodes to the unknown node to determine three circles. Based on the geometry theory, the point at which three circles intersect represents the coordinates of the unknown node. However, using three revised RSSs will still cause inevitable errors because we cannot completely eliminate the wall attenuation. Thus, we only utilize the RSSs from two beacon nodes to determine the position of the unknown node. The RSS from the third beacon node is not considered or it is only used as reference information. The analysis of the location of the two circles yielded five cases, as shown in Fig. 3.

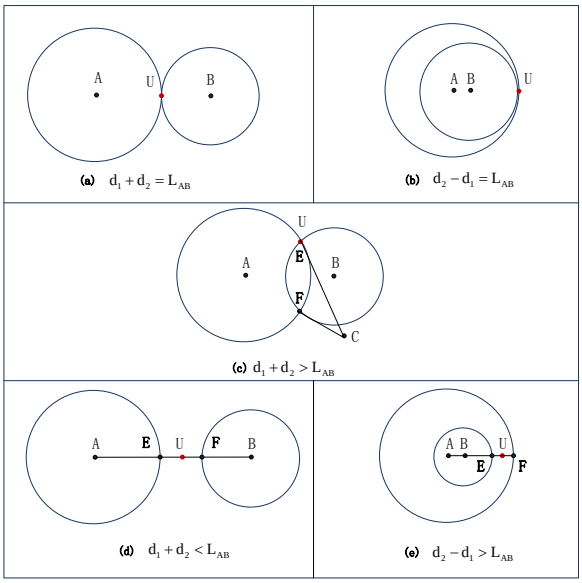


Figure 3. Five cases of using two circles to determine the location of the unknown node

When the two circles are externally tangent or internally tangent as shown in Figs. 3(a) and (b), the estimated intersection point of the unknown node is taken.

If the two circles intersect, there are two intersection points as shown in Fig. 3(c). In this case, we use a revised RSS between the third beacon node and the unknown node as the reference to compare $|d_3 - CE|$ with $|d_3 - CF|$ and select the smaller one as the estimated unknown node. Here, RSS from the third beacon node is utilized as a reference to identify the proper position between two intersection points.

In addition, as shown in Figs. 3(d) and (e), there may be no intersection point because of measurement error. In these cases, we take the middle point of E and F as the unknown node.

With respect to this improved method, we use the RSSs of two beacon nodes instead of three to decrease the error caused by the wall attenuation. However, the best case is when two RSSs without any wall attenuation are available, or that there are two zeros in some rows of the wall number matrix.

VI. SIMULATION RESULTS

In this section, the performances of the proposed localization and attenuation methods along with the improved method are demonstrated through computer simulation in MATLAB. As shown in Fig. 2, four beacon nodes are allocated in each floor whose area is 10 m×10 m, and the device in any position of this area is assumed to be able to detect all the beacon nodes. As there are two parallel walls, each horizontal plane is divided into three small areas *A*, *B*, and *C*. The corresponding wall number is shown in each row of the matrix given in (12). Here, both the estimated attenuation-based and prior attenuation-based revisions are applied to each area. Additional parameters are as follows: $PL(d_0)$ was set to 20 dB because free space propagation environment is assumed, and μ is set to be 2.0. The single wall attenuation and the floor attenuation were 20 dB and 30 dB, respectively. Here, Root Mean Square Error (RMSE) is defined in (15) to evaluate the localization performance.

$$RMSE = \sqrt{(x_{est} - x_{real})^2 + (y_{est} - y_{real})^2} \qquad (15)$$

where (x_{est}, y_{est}) and (x_{real}, y_{real}) are estimated coordinate and actual coordinate, respectively.

The effect of different noise variances on the localization performance for each area using two methods is simulated in Fig. 4. Through the simulation results, we observe that the accuracy of area A and C are nearly the same because these two areas suffer from similar wall attenuation. In addition, area B achieves the best localization performance. There is no gap between the wall attenuation from the beacon nodes. Thus, there is a similar error from each beacon node. When we use (13) and (14) to locate the unknown node in area B, some parts of that same error will be eliminated automatically, thereby further increasing the localization accuracy.

 ISBN: 978-1-61208-279-0

On the other hand, if we use prior wall attenuation to revise the RSS, the same performance can be obtained, irrespective of the location of the unknown node. As prior attenuation approaches focused mainly on actual attenuation, the revised RSS will be extremely close to the RSS without any walls. Thus, the presence of the walls does not give rise to any errors. However, there is always a gap between the prior attenuation and the actual attenuation. In worse cases, the prior information is not available. Thus, in most cases, we can only make use of the estimated attenuation to locate the target node.

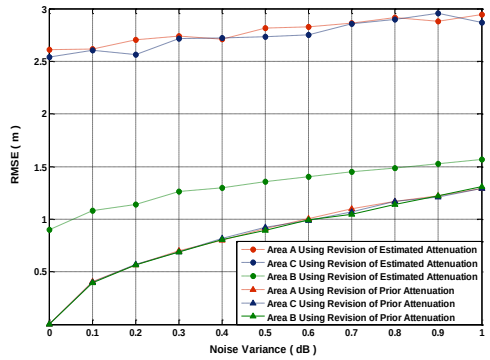


Figure 4. Localization performance in each area using two revision methods

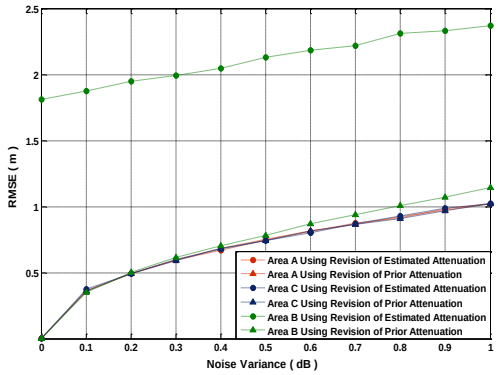


Figure 5. Localization performance in each area using two revision methods for improved algorithm

With respect to the improved method, using the same parameters as the previous method, we obtained the simulation performance in each area with two revision methods, as shown in Fig. 5. Obviously, the accuracy is increased, especially for areas *A* and *C* . However, we obtained a slightly worse performance in area *B* compared to the previous method. Some errors eliminated in area *B* in the previous method were not eliminated in this improved method. Hence, the previous method is the optimal choice for areas that suffer from the same wall attenuation from the beacon nodes. To this end, additional exploration is needed to determine the optimal choice between the two approaches used in this paper according to the detailed attenuation situation of the unknown node. And the performance is also influenced by *FA* and

VII. CONCLUSION

In this paper, we proposed a floor-wall attenuation estimation scheme, and exploited it to identify the floor number and to estimate the horizontal position of an unknown node. It is suitable for indoor environments with multiple floors and multiple walls. From the simulation results, the scheme achieved good accuracy, especially for the positioning of the planes. Based on the results, an improved method was utilized to increase the localization accuracy in parts of indoor areas. This method can work more efficiently if *a priori* information is available on interior structures and wall distribution.

ACKNOWLEDGMENT

This research was supported by the MSIP, Korea under the Convergence-ITRC support program (NIPA-2013-H0401-13-1004) supervised by the NIPA.

REFERENCES

[1] J. Hightower and G. Borriello, “Location systems for ubiquitous computing,” Computer, vol. 34, no. 8, Aug. 2001, pp. 57-66, doi:10.1109/2.940014.

[2] K. Pajlavan, X. Li, and J. Makela, “Indoor geolocation science and technology,” IEEE Commun Mag, vol. 40, no. 2, Feb. 2002, pp. 112-118, doi:10.1109/35.983917.

[3] F. Osterlind, E. Pramsten, D. Rpberthson, J. Eroksson, N. Finne, and T. Voigt, “Integrating building automation systems and wireless sensor networks,” Proc. IEEE ETFA 2007, Patras. Greece, Sept. 2007, pp. 1376-1379, doi:10.1109/EFTA.2007.4416941.

[4] C.-H. Chang, “Revisiting relative location estimation in wireless sensor networks,” Proc. IEEE ICC 2009, Dresdden. Germany, June. 2009, pp. 1-5, doi:10.1109/ICC.2009.5199419.

[5] J.-P. Sheu, “A distributed localization scheme for wireless sensor networks with improved grid-scan and vector-based refinement,” IEEE Trans. Mobile Comp., vol. 7, Sept. 2008, pp. 1110-1123, doi:10.1109/TMC.2008.35.

[6] A. Borrelli, C. Monti, M. Vari, and F. Mazzenga, “Channel models for IEEE 802.11b indoor system design,” Proc. IEEE ICC 2004, vol. 6, June 2004, pp. 3701-3705, doi: 10.1109/ICC.2004.1313233.

[7] S. Ranvier, “Path loss models,”Helsinki University of Technology, Helsinki, Finland, Nov. 2004 (Available online at http://www.comlab.hut.fi/opetus/333/2004_2005_slides/Path_loss_models.pdf).

[8] C. Y. Shih and P. J. Marron, “COLA: Complexity-reduced trilateration approach for 3D localization in wireless sensor networks,” Proc. IEEE SensorComm 2010, Venice. Italy, July. 2010, pp. 24-32, doi: 10.1109/SENSORCOMM.2010.11.

[9] R. Bolla, R. Rapuzzi, M. Repetto, P. Barsocchi, S. Chessa and S. Lenzi, “Automatic multimedia session migration by means of a context-aware mobility framework,” Proc. ICMA, Nice. Fance, Sep. 2009, doi: 10.1145/1710035.1710071.

[10] P. Barsocchi, S. Lenzi, S. Chessa and F. Furfari, “Automatic virtual calibration of range-based indoor localization systems,” Wirel. Commun. Mob. Comput, vol. 12, Feb. 2011, pp.1546-1557, doi: 10.1002/wcm.1085.

 ISBN: 978-1-61208-279-0

Carrier-Grade Internet Access Sharing in Wireless Mesh Networks: the Vision of the CARMNET Project

Mariusz Glabowski
Chair of Communication and Computer Networks
Poznan University of Technology
Poznan, Poland
e-mail: mariusz.glabowski@put.poznan.pl

Andrzej Szwabe
Institute of Control and Information Engineering
Poznan University of Technology
Poznan, Poland
e-mail: Andrzej.Szwabe@put.poznan.pl

Abstract — **The paper presents the vision of CARMNET – a Swiss-Polish project aimed at investigating "CARrier-grade delay-aware resource management for wireless multi-hop/Mesh NETworks". The project focuses on developing solutions that will motivate telecom operators to reconsider their view on user-operated IEEE 802.11-compliant wireless mesh networks. The project is driven by the vision of networks operated jointly by telecom operators – likely appreciating the CARMNET compliance with their IP Multimedia Subsystem (IMS) infrastructure – and a community of users contributing to and enjoying the pervasiveness of the CARMNET-based Internet access. The project aims at providing, both telecom operators and potential end users, with solutions that will create appropriately strong incentives – technological, functional and economical – for a widespread adoption of CARMNET-like networks within a steadily expanding group of users. Initial results indicate, that despite the originality of the project vision, the preliminary CARMNET system architecture complies with key relevant standards.**

Keywords – wireless mesh networks, user-operated Internet access sharing, IMS, AAA, NUM

I. INTRODUCTION

The core idea of the CARMNET project [1] is to make the user-provided Internet access an important alternative to the currently widespread 3G/4G-based mobile Internet access, in particular this provided in the femtocell scenario [2]. The main assumption of the project is that wireless mesh networks [3], while effectively enhanced by the introduction of advanced resource management mechanisms [4][5] and the compliance with the core of the telecom operators' IMS-based Authentication, Authorization, Accounting (AAA) infrastructure [6], may serve as an appropriate basis for a real-world realization of the core CARMNET idea. However, the successful realization of the vision of CARMNET networks – operated jointly by telecom operators and an informal community of Internet access-sharing users – requires facing several scientific and technological challenges that have not been yet completely investigated in the literature [7]. The assumed research tasks have to lead to determination of the comprehensive solution ensuring satisfactory levels of reliability and sustainability of the user-provided Internet access sharing [8]. The research will focus mostly on elaboration of algorithms for reliable servicing of multi-service traffic, with different packet delay tolerance, including algorithms related to: traffic stream classification, packet scheduling, buffer memory management, routing and nodes mobility management.

Targeting the CARMNET objectives implies the need for facing several technological challenges, in particular those related to the compatibility with the key relevant standards, such as Optimized Link State Routing (OLSR) protocol for the reliable multi-criteria routing within wireless mesh networks, or relevant to IMS-based AAA [6] technologies used by telecom operators. Moreover, as far as the long-term sustainability of CARMNET is concerned, some user-centric features are of the key importance, as well. They correspond to functional aspects of a CARMNET network use, such as the user-perceived network utility [5] and the user-friendliness of mobile applications running on smartphones that constitute such a network.

The further part of the paper is organized as follows. In Section II, the project's research motivation is presented. Section III describes the main phases of the project. In Section IV, the exemplary scenarios for CARMNET-like networks are presented and the research areas of the project are outlined. Section V concludes the paper.

II. RESEARCH MOTIVATION

The attractiveness of wireless mesh networks to telecom operators, despite a significant research effort that has been put in the last decade [3], remains quite limited. The following issues related to the CARMNET vision may be recognized as potentially postponing the wide adoption of existing wireless networking solutions:

- The lack of integration between the wireless network resource management and the AAA mechanism of telecom operators IMS-compliant networks,
- The lack of carrier-grade systems enabling telecom operators to measure the usage of shared Internet access in wireless mesh networks,
- The lack of solutions enabling end users to request the same level of Quality of Service (QoS) parameters as in 3G/4G networks,
- The lack of integration of the wireless network resource management oriented on the Network Utility

 ISBN: 978-1-61208-279-0

Maximization (NUM) with 'utility-aware' accounting, in particular in a scenario providing users with 'society-building' incentives similar to those familiar to users of popular Internet file-sharing applications based on the Peer-to-Peer (P2P) protocols [9].

It is worth mentioning that the scope of CARMNET research corresponds to the recent trend of intensive studies on various wireless Internet access sharing methods [7]. The need for CARMNET-like solutions may also be observed in efforts of several commercial Internet service providers, such as FON or Meraki [8]. Moreover, similar scientific projects have been recently conducted, including EU CARMEN [16]. However, to the best of our knowledge, all such initiatives differ from CARMNET in one of its core assumptions: they are based on the use of non-standard hardware.

III. CARMNET WORKPLAN

The CARMNET work plan is divided into three phases. The main goal of the CARMNET theoretical research (Phase 1) is to develop a framework and implementable solutions (i.e., the network architecture, resource management models and algorithms, as well as extensions of existing network protocols) enabling realization of the carrier-grade wireless mesh networks.

During the second phase, the theoretical results will be converted into technically implementable and commercially feasible network protocols and systems, which will be experimentally evaluated in realistic wireless testbeds.

Finally, the key implementation-oriented project outcome will be provided: an IMS-compliant prototype of a wireless network resource management system enabling the realization of 'charging per utility'. This way we hope to propose a solution capable to provide incentives for market-like, self-optimization of the Internet access sharing provided within a community of users.

Selected research topics of the project, as well as the scenarios of CARMNET network usage that will be considered during realization phases of the CARMNET project, are described in Section IV.

IV. CARMNET VISION

The implementation of the CARMNET vision, i.e., a wireless network that allows its end-user to share their network resources, requires defining the usage scenarios and the studies on several research areas implied by such definitions.

A. Scenarios

The CARMNET research methodology follows the approach of the user-centered scenario-based design, focused on functional specification of the system in correspondence to the user requirements and activities [10]. Since various wireless network topologies impose different user roles and activities, the main classification of CARMNET scenarios is made according to the types of connections (one-hop versus multi-hop) available in the wireless mesh network. The first scenario represents the case of the fully-connected mesh network (Figure 1), whereas the second scenario allows utilization of multi-hop mesh connections (Figure 2) and imposes the additional role of CARMNET relaying nodes. The figures show the components of the IMS architecture used in CARMNET: Home Subscriber Server (HSS), Proxy – Call Session Control Server (P-CSCF), and Serving-CSFS (S-CSCF), as well as Session Initiation Protocol (SIP) Servlet located at Application Server (AS) [6].

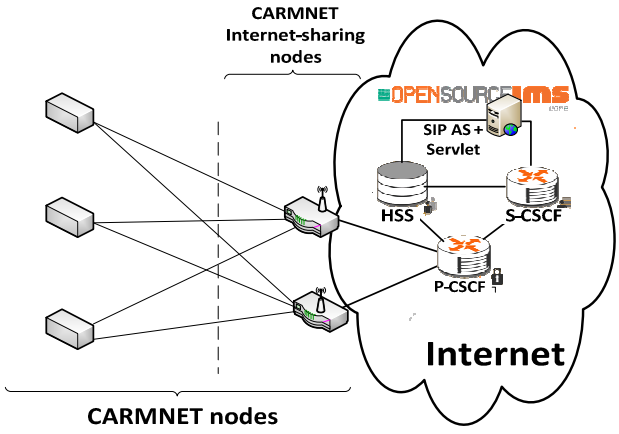


Figure 1. The CARMNET fully-connected mesh network scenario.

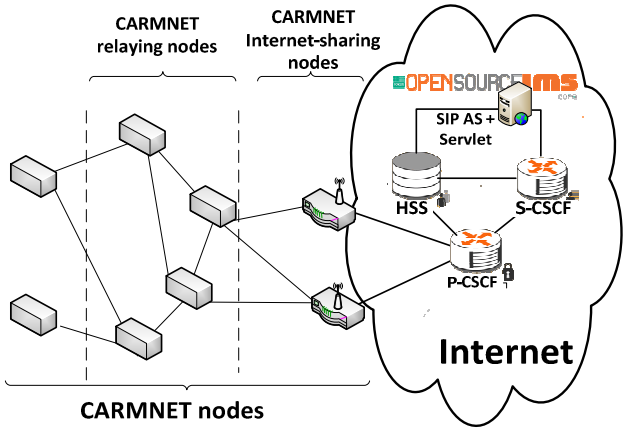


Figure 2. The CARMNET multi-hop mesh network scenario.

In both the CARMNET scenarios the basic network connectivity role is played by the CARMNET Internet sharing nodes, i.e., the network nodes, which may offer the Internet connection to users of other nodes.

B. Multi-Criteria Routing

Within the activities related to CARMNET project a new routing protocol (based on the OLSR protocol) that allows for multi-criteria path selection will be proposed. The

 ISBN: 978-1-61208-279-0

protocol will be capable to build the routing table (at each node), including not only the best path but a set of paths that lead to the specified destination network. The paths in the set will be selected as the subsequent shortest paths to the specified destination, based on one of the *k*-shortest paths algorithms [11][12][13]. The paths will be determined according to main criterion, e.g., delay, and they will include additional criteria (metrics). The additional metrics will be useful in order to choose the best path, that fulfills the criteria for a given traffic stream. The criteria will correspond to the QoS requirements for all traffic classes offered in the CARMNET network. An example of the criteria can be delay, a number of hops, link reliability or link load. Thus, the proposed QoS routing protocol will be able to use different traffic profiles and for each of them will propose the best path, i.e., the path, which fulfills recommended (for the considered traffic profile) QoS values in the best possible way.

The multi-criteria routing is dedicated primarily to the multi-hop scenario, but it can be also used in the single-hop scenario, to select the best CARMNET Internet-sharing node. Additionally, the routing protocol introduced in the CARMNET may be also used as one of the possible methods for mobility management: one or more of the criteria can be used by a mobile node to select the best path (next-hop node) for a traffic stream of a given class.

C. Utility-based Charging

The original CARMNET concept of the utility-based charging is largely based on a synergic combination of the following conceptual components (Figure 3):

- Charging per traffic volume,
- Traffic volume virtualization based on the mechanism of explicit transfer of virtual units that has been proposed as the key element of the Delay-Aware NUM System (DANUM) framework [5],
- IMS-based AAA, realized in the scenario of charging per traffic volume.

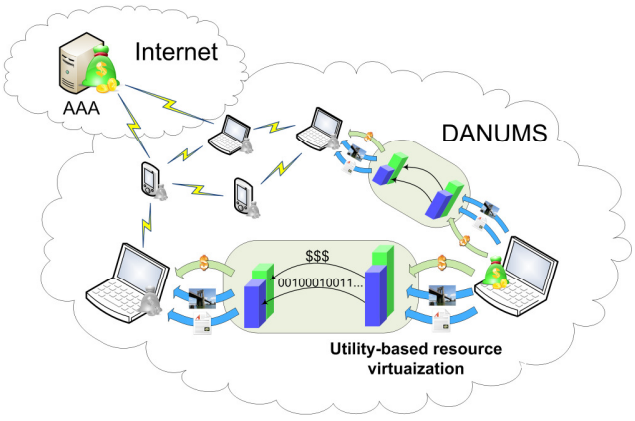


Figure 3. The concept of the utility-based charging as the integration of a DANUM system with an IMS-compliant AAA system.

Typically, CARMNET users are the ones that do not have the direct and acceptably cheap access to 3G/4G network. It is assumed that virtual utility units, after being earned by users sharing their mobile Internet access with other users of CARMNET-based wireless networks, may be spent by these users ('potentially altruistic', i.e., risking the lack of a reward for sharing the Internet connection) for accessing mobile Internet connection shared by other users.

D. Resource Management

One of the key objectives of CARMNET is to integrate IMS-based AAA support with the utility-oriented resource management for wireless mesh networks, in particular the one based on DANUM System (DANUMS) [5][4]. DANUMS is an application-layer system providing a delay-aware indirect flow control mechanism based on a system transporting virtual utility units and a packet forwarding component aimed at providing an approximation of Max-Weight Scheduling (MWS) [14]. The system is the first delay-aware NUM solution interoperable with widely used protocols such as Transmission Control Protocol (TCP), User Datagram Protocol (UDP), Internet Protocol (IP), and 802.11 Media Access Control (MAC) [5].

The other key objective of the CARMNET project in the area of resource management is to elaborate a model of a multiservice state-dependent queuing system with limited queue and state-dependent dynamic resource sharing between individual classes of calls. The advantage of the proposed model will be the possibility to evaluate analytically the average parameters of queues for individual classes of calls, which may prove to be of particular importance in engineering applications, especially in solutions concerning the analysis, dimensioning and optimization of mobile networks.

E. IMS-Compliant AAA Support

The CARMNET architecture assumes the application of an open implementation of the IMS server infrastructure (OpenIMS), extended by the SIP servlet located on the Application Server (AS). The communication between the network nodes and the IMS Server is realized with the use of CARMNET User Agent (a lean SIP client application).

It should be stressed that according to the vision of CARMNET, the standard session management functionalities provided by IMS core servers are used in a non-typical way - for the management of user-shared Internet access sessions (so called "CARMNET sessions") rather than, e.g., for the management of VoIP sessions. On the other hand, the standard AAA functionalities provided by IMS core are extended by additional CARMNET-specific features of utility monitoring that enable utility-based charging. These additional functionalities are provided in an IMS-complaint way, as a result of an implementation of SIP servlet and a special "CARMNET over SIP protocol" used for exchanging the information for the purpose of the utility-based charging. What is specific

 ISBN: 978-1-61208-279-0

for CARMNET is that the IMS infrastructure is used to manage users profiles and to store the configuration of end-users' utility functions.

F. Multi-Testbed Experimentation

The project aims at achieving a significant experimentally evaluated improvement in wireless network resource management. The experiment evaluation will involve both the effectiveness of network resource management systems and user-oriented network reliability.

In order to make the experimentation more reliable and effective, a special evaluation methodology and experiment description framework will be defined. A set of software components enabling an automated, highly controllable experimentation evaluation of IMS applications realized in wireless multi-hop networks will be developed.

The project outcomes will be evaluated in several experimental scenarios of infotainment and conversational services for networks of wire-line infrastructure limited to Internet access points, in wireless fully-connected mesh network and multi-hop network scenarios, as well as in mobility-oriented scenarios.

CARMNET solutions will be tested in several realistic testbeds, in particular in the ones located at CARMNET partners facilities (including the wnPUT testbed [4], and the SUPSI [1] testbed). Additionally, experiments in large-scale wireless testbeds, i.e., DES-Testbed [15] and NITOS testbed [4], are planned, as well as experiments performed within the facilities of a telecom operator.

V. CONCLUSIONS

In our opinion, CARMNET is a project worth a significant interest of researchers working in – so far rather distinct – areas of wireless mesh networking and IMS-based session and user management. On the other hand, the practical importance of the project research objectives seems to be in line with the recent trend of deploying wireless Internet access sharing by commercial service providers and telecom operators. The project is aimed at ensuring incentives for both telecom operators and potential mobile Internet end users for a fast and widespread adoption of CARMNET-like networks.

Although, for the time of writing this paper, the project is at an early stage, its initial results are already encouraging. In particular, despite the originality of the functional and the technological assumptions with regard to the CARMNET system, the initial system architecture design seems not to compromise the compliance with key relevant standards, such as OLSR, and IMS core standards.

We believe that, largely thanks to the originality of the CARMNET vision and the appropriateness of the experimentation-oriented methodology, CARMNET will provide outcomes of both the purely scientific impact and the practical added value.

ACKNOWLEDGMENT

This work was supported by a grant CARMNET financed under the Polish-Swiss Research Programme by Switzerland through the Swiss Contribution to the enlarged European Union.

REFERENCES

[1] CARrier-grade delay-aware resource management for wireless multi-hop/Mesh NETworks, http://www.carmnet.eu [retrieved: April, 2013].

[2] V. Chandrasekhar, J. Andrew, and A. Gatherer, "Femtocell networks: A survey," IEEE Commu. Magazine, vol. 46, IsFU 9, Sep. 2008, pp. 59-67.

[3] I. Akyildiz, X. Wang, and W. Wang, "Wireless mesh networks: a survey," Computer Networks 47 (4), 2005, pp. 445–487.

[4] K. Choumas, et al., "Optimization driven Multi-Hop Network Design and Experimentation: The Approach of the FP7 Project OPNEX," IEEE Communications Magazine, vol. 50, no. 6, June 2012, pp. 122-130.

[5] A. Szwabe, P. Misiorek, and P. Walkowiak, "Delay-Aware NUM system for wireless multi-hop networks," Proc. of 17th IEEE European Wireless 2011, EW2011, Vienna, Austria, April 27-29, 2011, pp. 530–537 .

[6] H. Khartabil, A. Niemi, M. Poikselka, and G. Mayer, "The IMS: IP multi-media concepts and services in the mobile domain. In The IMS: IP Multimedia Concepts and Services in the Mobile Domain", 2004, pp. 32–148.

[7] S. Jakubczak, D.G. Andersen, M. Kaminsky, K. Papagiannaki, and S. Seshan, "Link-alike: using wireless to share network resources in a neighborhood," ACM SIGMOBILE Mobile Computing and Communications Review, vol. 12 n.4, Oct. 2008, pp. 1-14.

[8] C. Middleton and A. Potter, "Is it good to share? A case study of FON and Meraki approaches to broadband provision," Proceedings of International Telecommunications Society 17th Biennial Conference, Montreal, 2008.

[9] M. Chen, M. Ponec, S. Sengupta, J. Li, and P. A. Chou, "Utility maximization in peer-to-peer systems," Proceedings of the 2008 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, June 02-06, 2008, Annapolis, MD, USA.

[10] J.M. Carroll and R.H.J. Sprague, "Five Reasons for Scenario-Based Design," Proc. of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32., 1999, pp. 3051-3061.

[11] A.W. Brander and M.C. Sinclair, "A Comparative Study of k-Shortest Path Algorithms," Proc. of 11th UK Performance Engineering Workshop, 1995, pp.370-379.

[12] D. Eppstein, "Finding the k Shortest Paths," SIAM Journal on Computing, vol. 28, no. 2, 1998, pp. 652-673.

[13] C.N. Clímaco, M.B. Pascoal, M.F. Craveirinha, M. Eugénia, and V. Captivo, "Internet packet routing: Application of a K-quickest path algorithm," European Journal of Operational Research, vol. 181, no. 3, Sep. 2007, pp.1045–1054.

[14] L. Georgiadis, M. J. Neely, and L. Tassiulas. "Resource allocation and cross-layer control in wireless networks." Foundations and Trends in Networking, 2006, pp. 1–149.

[15] A. Szwabe, P. Misiorek, M. Urbański, F. Juraschek, and M. Guenes : "Multi-path OLSR Performance Analysis in a Large Testbed Environment," in L. Bononi et al. (Eds.): ICDCN 2012, LNCS, vol. 7129, Springer, 2012, pp. 488–501.

[16] A. Banchs, et al., "CARMEN: Delivering Carrier Grade Services over Wireless Mesh Networks" PIMRC 2008, 15-18 Sep. 2008, pp. 1-6.

 ISBN: 978-1-61208-279-0

Quantization Errors in Overlapped Block Digital Filtering Methods

Mustafa Daloglu
Dept. of Electrical and Electronics Eng.
Bilkent University
Bilkent, Ankara, Turkey
email: mudaloglu@gmail.com

Erchin Serpedin
Dept. of Electrical and Computer Eng.
Texas A&M University
College Station, Texas, USA
email: serpedin@ece.tamu.edu

***Abstract*—In digital signal processing applications involving filtering long sequences, block filtering methods like overlap-save and overlap-add are widely used. Like all finite-precision applications, overlap-save and overlap-add methods are also affected by quantization errors. The goal of this paper is to calculate and make a quantitative comparison of the overall quantization noise resulting from the two methods in terms of power (variance) of the quantization noise. Multiple quantization noise sources are taken into consideration in the computation of the variances. The calculations reveal that the overlap-add approach is more prone to quantization noise compared to the overlap-save approach due to the addition of overlapping sections between overlap-add output blocks.**

Keywords-block filter; quantization noise; overlap-add; overlap-save;

I. INTRODUCTION

Overlap-save and overlap-add block filtering algorithms are frequently used to implement block finite impulse response (FIR) filters [1], particularly when the input sequences are long. In these widely used approaches, segments from the long input signal are processed using the discrete Fourier transform (DFT), which is generally implemented via the fast Fourier transform (FFT) [2]. In all the computational stages of the block filtering approaches, quantization errors occur due to the necessity of expressing numerical quantities with finite precision in digital signal processors. The goal of this study is to estimate the overall quantization noise in the output blocks resulting from the overlap-save (OLS) and overlap-add (OLA) methods, and compare these errors in terms of quantization noise power. In order to compute the overall quantization noise in the output blocks, multiple error sources related to the different stages of the computation must be considered, as for example, it is the case with the errors resulting from the analog-to-digital (A/D) conversion and the DFT computation stage.

Individual quantization error sources have been amply investigated in previous studies. These include the quantization errors from the A/D conversion process [3,4], errors generated from a standard DFT computation [3,4] and quantization errors from an FFT process [3,4,5]. However, there is no previous attempt to investigate the overall quantization error present in the output blocks of the OLA approach and OLS scheme, to the best of our knowledge. In the following sections, we will try to estimate and compare the power of the overall quantization error in the output blocks of the OLA and OLS processes computed via the standard DFT.

The rest of this paper is organized as follows. In Section II, the probabilistic properties of the quantization noise generated in the different stages of the computational process are described. In Section III, the variance of the quantization error present in the output of the OLS blocks is calculated. In Section IV, the same calculation is conducted for the output blocks of the OLA. A brief numerical study is conducted to illustrate the difference in terms of quantization noise power between OLA and OLS methods. In Section V, the error variances from the two different OLA and OLS block filters are compared and several concluding remarks are presented.

Throughout the paper, the DFT and the inverse DFT (IDFT) of any signal block are computed via multiplication with the DFT and IDFT matrix, respectively. The $N \times N$ DFT matrix for an N-point DFT computation is defined as [6]:

$$\boldsymbol{W_N} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & e^{-j\frac{2\pi}{N}} & \dots & e^{-j\frac{2\pi(N-1)}{N}} \\ \vdots & \vdots & & \vdots \\ 1 & e^{-j\frac{2\pi(N-1)}{N}} & \dots & e^{-j\frac{2\pi(N-1)^2}{N}} \end{bmatrix}. \quad (1)$$

The $N \times N$ IDFT matrix for an N-point IDFT computation is given by [6]:

$$\boldsymbol{W_N^{-1}} = \frac{1}{N}\boldsymbol{W_N^H}, \quad (2)$$

where H stands for the complex conjugate transpose (Hermitian). In addition, $\boldsymbol{x}$ denotes the time-domain input sequence and $\widehat{\boldsymbol{x}}$ represents the DFT of $\boldsymbol{x}$, and it is considered to belong to the frequency domain. This relationship can be expressed compactly as:

$$\widehat{\boldsymbol{x}} = \boldsymbol{W_N}\boldsymbol{x}. \quad (3)$$

Similarly, $\boldsymbol{h}$ denotes the time-domain impulse response of the filter and $\widehat{\boldsymbol{h}}$ represents the DFT of $\boldsymbol{h}$.

II. PROBABILISTIC PROPERTIES OF THE QUANTIZATION NOISE

In this section, we describe from a statistical perspective the quantization noises generated in the different stages of

 ISBN: 978-1-61208-279-0

the OLA and OLS methods. The quantization noise in all of these stages is modeled as an additive noise (or error) signal [4, p.119]. A statistical model is used for this additive noise with the assumptions that: "the error sequence $\boldsymbol{e}$ is a sample sequence of a stationary random process, $\boldsymbol{e}$ is uncorrelated with the original signal, the random variables (RV) of the error process are uncorrelated and the probability distribution functions (PDF) of the error process are uniform over the range of the quantization error [4, p.120]." For quantizers using rounding in quantization, the amplitude of $\boldsymbol{e}$ is in the range:

$$\frac{-\Delta}{2} < \boldsymbol{e} < \frac{\Delta}{2}. \tag{4}$$

With the assumptions above, the mean value of $\boldsymbol{e}$ is zero and its variance (power) is [4, p.120]:

$$\sigma_e^2 = \frac{\Delta^2}{12}. \tag{5}$$

A. Quantization Noise in the A/D Conversion

Quantization noise occurs at the A/D conversion stage, where the analog-to-digital converter (ADC) is the quantizer. Referring to the noise model above, the step size Δ_{adc} of a $B+1$-bit two's complement ADC is in general given by [4, p.118]:

$$\Delta_{adc} = \frac{2X_m}{2^{B+1}} = \frac{X_m}{2^B}, \tag{6}$$

where X_m is called the full-scale level of the ADC [4, p.118]. Consequently, the variance of the ADC quantization noise σ_{adc}^2 is expressed as:

$$\sigma_{adc}^2 = \frac{\Delta_{adc}^2}{12} = \frac{X_m^2}{3}2^{-(2B+2)}. \tag{7}$$

B. Quantization of the Filter Coefficients

The filter coefficients, or the impulse response $\boldsymbol{h}$, are stored in a finite-precision digital environment in digital filtering applications which also require quantization. If the impulse response coefficients are rounded to $B+1$ bits, then the step size of this quantization Δ_h will be [4,p.347], [3,p.580]:

$$\Delta_h = 2^{-B}. \tag{8}$$

Then, the variance of this noise σ_h^2 will be:

$$\sigma_h^2 = \frac{\Delta_h^2}{12} = \frac{2^{-(2B+2)}}{3}. \tag{9}$$

C. Quantization Noise in an N-Point DFT

The DFT operation is also implemented in a finite-precision digital environment, for this reason quantization noise is also present in the computation of a DFT. The main source of the error is represented by the round-off errors due to the multiplications performed in the DFT with fixed point arithmetic; therefore, we assume that the DFT coefficients themselves and the addition operations are not quantized. Assuming that the input signal is also complex valued, there are N complex-valued multiplications (thus $4N$ real-valued multiplications) in the direct computation of the DFT. We also assume that these $4N$ quantization errors are mutually uncorrelated and also uncorrelated with the input sequence [3, p.487]. If the numbers are represented with $B+1$ bits, the variance of the error resulting form a single real multiplication would be $2^{-2B}/12$ [4, p.630]. Since we have $4N$ mutually uncorrelated errors, the variance of the overall error would be expressed as [4, p.631], [3, p.487]:

$$\sigma_{dft}^2 = \frac{N}{3}2^{-2B}. \tag{10}$$

D. Quantization Noise in the Multiplication $\widehat{\boldsymbol{x}} \odot \widehat{\boldsymbol{h}}$

After computing the DFTs of the input signal and the impulse response, the next step is to multiply the two N-point frequency domain sequences term by term using the Hadamard product "$\odot$". This operation corresponds to a circular convolution in the time-domain and it is actually representing the filtering process. For each term, there will be a single complex-valued multiplication or 4 real-valued multiplications. Using the same approach as above, we can directly state that the error corresponding to this operation (supposing that the representation assumes $B+1$ bits) will have a variance of:

$$\sigma_{mult}^2 = \frac{4}{12}2^{-2B} = \frac{1}{3}2^{-2B}. \tag{11}$$

E. Quantization Noise in an N-Point IDFT

The IDFT operation is symmetric to the DFT operation. Therefore, the previous analysis conducted for the DFT is directly extendable to the IDFT operation. The variance of the error generated by the IDFT operation will be given by:

$$\sigma_{idft}^2 = \frac{N}{3}2^{-2B}. \tag{12}$$

III. Quantization Noise in the OLS Method

One of the most widely used approaches for block digital filtering is the overlap and save method [7]. Considering an N-point input segment $\boldsymbol{x}$ and a P-point impulse response $\boldsymbol{h}$, the OLS method corresponds to implementing an N-point circular convolution and identifying the part of this convolution that corresponds to a linear convolution. Then, patching these segments together to form the output block. This can be obtained by dividing $\boldsymbol{x}$ into sections of length N such that each input section overlaps the preceding section by $P-1$ points, and by discarding the first $P-1$ samples of the output block that result from the N-point circular convolution. The remaining $N-(P-1)$ samples correspond to the linear convolution and can be patched block by block to form the output [4, p.558].

In order to implement the N-point circular convolution, we will first pad the impulse response with $N-P$ zeros so that it will have the same length as the input block (assuming that $N > P$). Then both of the time-domain

 ISBN: 978-1-61208-279-0

sequences will be transformed with the DFT operation and multiplied in the frequency-domain. The resulting sequence will be converted back to the time-domain via the IDFT operation. The quantization errors will be introduced step by step during the process and will be carried to the output. Any quantization error is denoted as an additive random vector with the same length as that of the original signal, i.e., it is modeled as a vector of uncorrelated complex random variables with a length N.

A. Computation of the Error Term

The first source of quantization error is the quantization of the input signal by the ADC and the quantization of the filter coefficients:

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{x_0} + \boldsymbol{\Delta x} \\ \boldsymbol{h} &= \boldsymbol{h_0} + \boldsymbol{\Delta h}, \end{aligned} \tag{13}$$

where $\boldsymbol{x_0}$, $\boldsymbol{h_0}$ are the original vectors and $\boldsymbol{\Delta x}$, $\boldsymbol{\Delta h}$ are the random vectors, all of length N. The next step is to transform both of these vectors to the frequency-domain via the multiplication with the N-point DFT matrix $\boldsymbol{W}$, which will also generate quantization errors:

$$\begin{aligned} \widehat{\boldsymbol{x}} &= \boldsymbol{W x_0} + \boldsymbol{W \Delta x} + \widehat{\boldsymbol{\delta x}} \\ &= \widehat{\boldsymbol{x_0}} + \widehat{\boldsymbol{\Delta x}} + \widehat{\boldsymbol{\delta x}}, \end{aligned} \tag{14}$$

where $\widehat{\boldsymbol{\delta x}}$ is the quantization noise generated from the DFT operation. In a similar fashion:

$$\begin{aligned} \widehat{\boldsymbol{h}} &= \boldsymbol{W h_0} + \boldsymbol{W \Delta h} + \widehat{\boldsymbol{\delta h}} \\ &= \widehat{\boldsymbol{h_0}} + \widehat{\boldsymbol{\Delta h}} + \widehat{\boldsymbol{\delta h}}. \end{aligned} \tag{15}$$

At this stage we multiply both vectors using the Hadamard product "$\odot$" in the frequency-domain to form the output:

$$\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{h}} \odot \widehat{\boldsymbol{x}} + \widehat{\boldsymbol{\mu y}}, \tag{16}$$

where $\widehat{\boldsymbol{\mu y}}$ is the quantization error vector generated from the finite-precision multiplications at this stage. After taking the Hadamard product of $\widehat{\boldsymbol{h}}$ and $\widehat{\boldsymbol{x}}$ and discarding the terms that result from the multiplication of two error terms, it follows that

$$\begin{aligned} \widehat{\boldsymbol{y}} = \widehat{\boldsymbol{h_0}} \odot \widehat{\boldsymbol{x_0}} + \widehat{\boldsymbol{h_0}} \odot \widehat{\boldsymbol{\Delta x}} + \widehat{\boldsymbol{h_0}} \odot \widehat{\boldsymbol{\delta x}} \\ + \widehat{\boldsymbol{\Delta h}} \odot \widehat{\boldsymbol{x_0}} + \widehat{\boldsymbol{\delta h}} \odot \widehat{\boldsymbol{x_0}} + \widehat{\boldsymbol{\mu y}}. \end{aligned} \tag{17}$$

The last quantization error source will be the one induced by the IDFT operation. We will model the noise induced by the IDFT in a similar way to that generated by the DFT process:

$$\boldsymbol{y} = \frac{1}{N} \boldsymbol{W^H} \widehat{\boldsymbol{y}} + \boldsymbol{\delta y}, \tag{18}$$

where $\boldsymbol{\delta y}$ stands for the quantization error vector generated by the IDFT operation.

In order to compute the power of the overall error, we express $\boldsymbol{y}$ in terms of the original signal and an error term:

$$\boldsymbol{y} = \boldsymbol{y_0} + \boldsymbol{\Delta y} \rightarrow \boldsymbol{y_0} = \frac{1}{N} \boldsymbol{W^H} (\widehat{\boldsymbol{h_0}} \odot \widehat{\boldsymbol{x_0}}). \tag{19}$$

Then we can express the final error in open form as:

$$\begin{aligned} \boldsymbol{\Delta y} = \frac{1}{N} \boldsymbol{W^H} (\widehat{\boldsymbol{h_0}} \odot \widehat{\boldsymbol{\Delta x}}) + \frac{1}{N} \boldsymbol{W^H} (\widehat{\boldsymbol{\Delta h}} \odot \widehat{\boldsymbol{x_0}}) \\ + \frac{1}{N} \boldsymbol{W^H} (\widehat{\boldsymbol{h_0}} \odot \widehat{\boldsymbol{\delta x}}) + \frac{1}{N} \boldsymbol{W^H} (\widehat{\boldsymbol{\delta h}} \odot \widehat{\boldsymbol{x_0}}) \\ + \frac{1}{N} \boldsymbol{W^H} (\widehat{\boldsymbol{\mu y}}) + \boldsymbol{\delta y}. \end{aligned} \tag{20}$$

As expressed above, the overall error is composed of 6 terms, which will be denoted respectively as $\boldsymbol{z_i}$, where $1 \leq i \leq 6$, for easy reference in the further stages:

$$\boldsymbol{\Delta y} = \sum_{i=1}^{6} \boldsymbol{z_i}. \tag{21}$$

The error vector $\boldsymbol{z_i}$ is an $N \times 1$ vector of complex random variables generated from the i^{th} quantization error source.

B. Variance of the Error Terms

The calculation of the output signal $\boldsymbol{y}$, showed that there are multiple quantization noise sources in the process, which are assumed to generate mutually uncorrelated and zero-mean noise vectors. For the computation of the variance, it is important to note that the noise vectors are mutually uncorrelated and their entries are also uncorrelated one with respect to the other. In other words, the noise terms in an individual vector are all uncorrelated with each other. With these assumptions, we will determine our strategy to calculate the variance of the overall noise term using $\boldsymbol{z}_{N \times 1}$ as a generic vector with a variance of σ_z^2. Note that $\boldsymbol{z}$ is a zero-mean noise vector:

$$E(\boldsymbol{z}) = \begin{bmatrix} E(z_1) \\ \vdots \\ E(z_N) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = \boldsymbol{0}. \tag{22}$$

Using this property and the fact that the noise terms are uncorrelated with each other, it follows that:

$$E(\boldsymbol{z z^H}) = \begin{bmatrix} E(|z_1|^2) & \dots & 0 \\ & \ddots & \\ 0 & \dots & E(|z_N|^2) \end{bmatrix} = \sigma_z^2 \boldsymbol{I}. \tag{23}$$

Note that σ_z^2 is the variance of a single term in the vector $\boldsymbol{z}$. In order to calculate the variance of the vector, we need to add up all the variances corresponding to all the entries in the vector. Mathematically, this operation amounts to

$$\sigma_{z-vector}^2 = E(\boldsymbol{z}^H \boldsymbol{z}) = N \sigma_z^2. \tag{24}$$

Once the variances of the individual noise sources are calculated, it is possible to find the overall variance just by adding the individual ones due to the uncorrelated nature

 ISBN: 978-1-61208-279-0

of the noise sources [8, p.220]. We will denote the variance of an individual noise vector $\boldsymbol{z_i}$ as σ_i^2, where $1 \leq i \leq 6$ denotes the respective order.

1) σ_1^2 and σ_2^2: Let's start with $\boldsymbol{z_1}$ and proceed using the strategy developed for the generic vector.

$$\begin{aligned} &\boldsymbol{z}_1 = \frac{1}{N}\boldsymbol{W^H}(\widehat{\boldsymbol{h_0}} \odot \widehat{\boldsymbol{\Delta x}}) \rightarrow E(\boldsymbol{z}_1^H \boldsymbol{z}_1) = \\ &E(\frac{1}{N^2}(\widehat{\boldsymbol{h_0}} \odot \widehat{\boldsymbol{\Delta x}})^H \boldsymbol{W}\boldsymbol{W^H}(\widehat{\boldsymbol{h_0}} \odot \widehat{\boldsymbol{\Delta x}})) \\ &= E(\frac{1}{N}(\widehat{\boldsymbol{h_0}} \odot \widehat{\boldsymbol{\Delta x}})^H (\widehat{\boldsymbol{h_0}} \odot \widehat{\boldsymbol{\Delta x}})), \end{aligned} \tag{25}$$

since $\boldsymbol{W}\boldsymbol{W^H} = N\boldsymbol{I}$. Writing in open form:

$$\begin{aligned} E(\boldsymbol{z}_1^H \boldsymbol{z}_1) &= E\big(\frac{1}{N}\sum_{i=1}^{N} |\widehat{h_0}[i]|^2 |\widehat{\Delta x}[i]|^2\big) \\ &= \frac{1}{N}\sum_{i=1}^{N} |\widehat{h_0}[i]|^2 E(|\widehat{\Delta x}[i]|^2), \end{aligned} \tag{26}$$

where $\widehat{\Delta x}$'s are the probabilistic terms. Note that $\widehat{\boldsymbol{\Delta x}} = \boldsymbol{W}\boldsymbol{\Delta x}$, and $\boldsymbol{\Delta x}$ is the zero-mean noise vector generated at the ADC composed of terms with a variance σ_{adc}^2 as defined previously. Therefore, the following relations hold:

$$\begin{aligned} &\widehat{\boldsymbol{\Delta x}} = \boldsymbol{W}\boldsymbol{\Delta x}\ ,\ E(\widehat{\boldsymbol{\Delta x}}) = 0 \\ &E(\widehat{\boldsymbol{\Delta x}}\widehat{\boldsymbol{\Delta x}}^H) = E(\boldsymbol{W}\boldsymbol{\Delta x}(\boldsymbol{W}\boldsymbol{\Delta x})^H) \\ &\qquad = \boldsymbol{W}E(\boldsymbol{\Delta x}\boldsymbol{\Delta x}^H)\boldsymbol{W^H} \\ &\qquad = \boldsymbol{W}\sigma_{adc}^2 \boldsymbol{I}\boldsymbol{W^H} = \sigma_{adc}^2 N\boldsymbol{I}. \end{aligned} \tag{27}$$

This means that each term in the vector $\widehat{\boldsymbol{\Delta x}}$ has a variance equal to $\sigma_{adc}^2 N$. Thus, we can replace any $E(|\widehat{\Delta x}[i]|^2)$ term with $\sigma_{adc}^2 N$. Therefore, it follows further that

$$\begin{aligned} E(\boldsymbol{z}_1^H \boldsymbol{z}_1) &= \frac{1}{N}\sum_{i=1}^{N} |\widehat{h_0}[i]|^2 \sigma_{adc}^2 N \\ &= \sigma_{adc}^2 \sum_{i=1}^{N} |\widehat{h_0}[i]|^2. \end{aligned} \tag{28}$$

Note also that $\widehat{\boldsymbol{h_0}} = \boldsymbol{W}\boldsymbol{h_0}$. Thus:

$$\begin{aligned} \sum_{i=1}^{N} |\widehat{h_0}[i]|^2 &= \widehat{\boldsymbol{h_0}}^H \widehat{\boldsymbol{h_0}} = (\boldsymbol{W}\boldsymbol{h_0})^H (\boldsymbol{W}\boldsymbol{h_0}) \\ &= \boldsymbol{h_0}^H \boldsymbol{W}^H \boldsymbol{W}\boldsymbol{h_0} = N\boldsymbol{h_0}^H \boldsymbol{h_0} \\ &= N\sum_{i=1}^{N} h_0^2[i]. \end{aligned} \tag{29}$$

To sum up, the variance (σ_1^2) of the quantization noise from the first source ($\boldsymbol{z_1}$) is found to be:

$$\sigma_1^2 = N\sigma_{adc}^2 \sum_{i=1}^{N} h_0^2[i]. \tag{30}$$

This also means that any term in the vector $\boldsymbol{z_1}$ has a variance equal to $\sigma_{adc}^2 \sum_{i=1}^{N} h_0^2[i]$.

We notice that the variance (σ_2^2) calculation of the quantization noise from the second source ($\boldsymbol{z_2}$) is similar with the first calculation. For this reason, we can directly state that:

$$\sigma_2^2 = N\sigma_h^2 \sum_{i=1}^{N} x_0^2[i], \tag{31}$$

where σ_h^2 is the noise caused by the quantization of the impulse response.

2) σ_3^2 and σ_4^2: The strategy for the calculation of these variances is the same as the one used for the first variance. The vector variance σ_3^2 of the noise generated from the third source $\boldsymbol{z_3}$ could be directly expressed as:

$$\sigma_3^2 = \frac{1}{N}\sum_{i=1}^{N} (|\widehat{h_0}[i]|^2 E(|\widehat{\delta x}[i]|^2). \tag{32}$$

As stated previously, $\widehat{\boldsymbol{\delta x}}$ is the zero-mean noise vector generated by the N-point DFT process with a variance σ_{dft}^2. Then, we can replace any $E(|\widehat{\delta x}[i]|^2)$ term with σ_{dft}^2. We have previously found out that $\sum_{i=1}^{N} |\widehat{h_0}[i]|^2 = N\sum_{i=1}^{N} h_0^2[i]$. Thus:

$$\sigma_3^2 = \sigma_{dft}^2 \sum_{i=1}^{N} h_0^2[i]. \tag{33}$$

As it was the case for the first set of variances, calculation of σ_3^2 and σ_4^2 is quite similar. Therefore, one can directly express σ_4^2 as follows:

$$\sigma_4^2 = \sigma_{dft}^2 \sum_{i=1}^{N} x_0^2[i]. \tag{34}$$

3) σ_5^2 and σ_6^2: The computation of these variances is less complex with respect to the previous variance calculations because $\boldsymbol{z_5}$ and $\boldsymbol{z_6}$ are generated at the final stages of the OLS process. For this reason, these noise vectors are carried through fewer computational stages in the OLS process with respect to the previous noise vectors. Calculation of $\boldsymbol{z_5}$ relies on the following relation:

$$\begin{aligned} \sigma_5^2 &= E([\frac{1}{N}\boldsymbol{W^H}\widehat{\boldsymbol{\mu y}}]^H [\frac{1}{N}\boldsymbol{W^H}\widehat{\boldsymbol{\mu y}}]) \\ &= \frac{1}{N}E(\widehat{\boldsymbol{\mu y}}^H \widehat{\boldsymbol{\mu y}}) = \frac{1}{N}\sum_{i=1}^{N} E(|\widehat{\mu y}[i]|^2). \end{aligned} \tag{35}$$

From our previous definitions, $\widehat{\boldsymbol{\mu y}}$ was the zero-mean noise vector generated by the Hadamard product of two complex N-point vectors in a finite precision environment with a variance σ_{mult}^2. Replacing $E(|\widehat{\mu y}[i]|^2)$ with σ_{mult}^2, we obtain:

$$\sigma_5^2 = \sigma_{mult}^2. \tag{36}$$

 ISBN: 978-1-61208-279-0

The last quantization noise source is the IDFT process, generating a noise vector $\boldsymbol{z_6} = \boldsymbol{\delta y}$ with terms having a variance σ^2_{idft}:

$$\sigma^2_6 = E(\boldsymbol{z}^H_6 \boldsymbol{z}_6) = \sum_{i=1}^{N} E(|\delta y[i]|^2). \tag{37}$$

The resulting value is:

$$\sigma^2_6 = N\sigma^2_{idft}. \tag{38}$$

Exploiting the assumption that all these 6 noise sources are mutually uncorrelated, we express the overall vector variance in an N-point output block computed via the OLS method as:

$$\begin{aligned}\sigma^2_{OLS} = (N\sigma^2_{adc} + \sigma^2_{dft}) \sum_{i=1}^{N} h^2_0[i] \\ + (N\sigma^2_h + \sigma^2_{dft}) \sum_{i=1}^{N} x^2_0[i] \\ + \sigma^2_{mult} + N\sigma^2_{idft}.\end{aligned} \tag{39}$$

All these variances were numerically determined in the second section. Considering a $(B+1)$-bit digital system that assumes a $(B+1)$-bit ADC with a full-scale level of X_m, we find the variance in an N-point OLS output block to be:

$$\begin{aligned}\sigma^2_{OLS} = \frac{N}{3} 2^{-2B} \big((\frac{X^2_m + 4}{4}) \sum_{i=1}^{N} h^2_0[i] \\ + (\frac{5}{4}) \sum_{i=1}^{N} x^2_0[i] + \frac{N^2+1}{N} \big).\end{aligned} \tag{40}$$

IV. Quantization Noise in the OLA Method

The other widely used block digital filtering method is the overlap and add method [1]. Considering an L-point input signal block $\boldsymbol{x}$ and a P-point filter impulse response $\boldsymbol{h}$, the OLA method convolves $\boldsymbol{x}$ and $\boldsymbol{h}$ using a linear convolution operation. Thus, it results in an output block of length $N = L+P-1$. In order to perform this operation in the frequency-domain via DFT-IDFT, we must pad both $\boldsymbol{x}$ and $\boldsymbol{h}$ with zeros such that both of them become vectors of the same length N, which corresponds to performing an N-point circular convolution. When patching the N-point output blocks together, the first $P-1$ terms of each block are overlapped and added with the last $P-1$ terms of the previous block [4, p.558].

In terms of variance calculation, an important difference between the OLS scheme and OLA method is that the OLA method requires the addition of the first $P-1$ terms of the current output block with the overlapping $P-1$ terms of the previous output block. This means that in these overlapping sections, the quantization errors present in the two individual blocks are also added. Other than this difference, the computation stages for both the OLS scheme and the OLA method are the same. Thus, we can use our previous variance calculations for the OLS and modify them appropriately to include the additional noise coming from overlapping $P-1$ terms.

Assume that the OLA method results in an output block of length N, and consider an N-point slice from the output sequence of the OLS. Just before overlapping and adding the the first $P-1$ terms of the current OLA block with the last $P-1$ terms of the previous block, both blocks from the OLA and the OLS process present the same variance. Assuming that the overlapping blocks of the OLA are uncorrelated, the $P-1$ noise terms being added to each other are also uncorrelated. It would be correct to state that the variance of the terms in the overlapping section of the OLA block will be doubled. With this reasoning, we can express the variance of the quantization noise in an N-point OLA output block (σ^2_{OLA}) in terms of the variance of the quantization noise in an N-point OLS output block (σ^2_{OLS}) as follows:

$$\sigma^2_{OLA} = \frac{N+P-1}{N} \sigma^2_{OLS}. \tag{41}$$

Adopting a $(B+1)$-bit digital system that assumes a $(B+1)$-bit ADC with a full-scale level of X_m, we can express the numerical value of σ^2_{OLA} as:

$$\begin{aligned}\sigma^2_{OLA} = \frac{N+P-1}{3} 2^{-2B} \\ \big((\frac{X^2_m + 4}{4}) \sum_{i=1}^{N} h^2_0[i] + (\frac{5}{4}) \sum_{i=1}^{N} x^2_0[i] + \frac{N^2+1}{N} \big).\end{aligned} \tag{42}$$

In practice, the difference in terms of quantization noise variance between OLS and OLA is relatively small. For example, when $N = 256$ and $P = 32$, according to (41), the difference is $10\log_{10}((N+P-1)/N) = 0.49dB$, while for $N = 256$ and $P = 128$, the difference amounts to $1.7dB$.

V. Conclusion

The quantization noise is a side effect that is inherently present in any digital system due to the necessity of representing the signal samples in finite-precision. Having an important role in digital signal processing applications that involve the filtering of long sequences, OLA and OLS methods are also prone to quantization noise once that they get implemented in a digital system. Both methods were implemented in multiple stages like the ADC, DFT or the IDFT stage and multiple quantization noise sources from these stages were taken into consideration. It was observed that the power of the quantization noise in an OLA output block tended to be higher than the quantization noise in an OLS output block, mainly because of the addition of the overlapping sections in the OLA process. However, it was also observed that the difference in the quantization noise power was highly dependent on the length of the input sequences.

 ISBN: 978-1-61208-279-0

The computation of the variances would have been more realistic if an FFT algorithm was used instead of a direct DFT computation because most digital signal processing applications take advantage of FFT algorithms. This study can be expanded to include an FFT algorithm instead of a direct DFT implementation. However, it is expected that the quantitative comparison of the quantization noises in OLA and OLS will not change.

ACKNOWLEDGMENT

This work was made possible by the support offered by QNRF-NPRP grants 09-341-2128 and 4-1293-2-513.

REFERENCES

[1] M. J. Narasimha, "Linear convolution using skew–cyclic convolutions," IEEE Signal Processing Letters, vol. 14, 2007, pp. 173–176.

[2] A. Daher, E. H. Baghious, G. Burel, and E. Radoi, "Overlap–save and overlap–add filters: optimal design and comparison," IEEE Transactions on Signal Processing, vol. 58, 2010, pp. 3066–3075.

[3] J. G. Proakis and D. G. Manolakis, Digital Signal Processing: Principles, Algorithms and Applications. Upper Saddle River: Prentice Hall, 1996.

[4] A. V. Oppenheim and R. W. Schafer, Discrete Time Signal Processing. Englewood Cliffs: Prentice Hall, 1989.

[5] A. V. Oppenheim and C. J. Weinstein, "Effects of finite register length in digital filtering and the fast Fourier transform," Proceedings of IEEE, vol. 60, 1972, pp. 957–976.

[6] A. E. Cetin, Lecture Notes on Discrete-Time Signal Processing: EE424 Course at Bilkent University, Ankara, Turkey, 2012.

[7] G. Burel, "Optimal design of transform-based block digital filters using a quadratic criterion," IEEE Transactions on Signal Processing, vol. 52, 2004, pp. 1964–1974.

[8] D. P. Bertsekas and J. N. Tsitsiklis, Introduction to Probability. Belmont: Athena Scientific, 2008.

 ISBN: 978-1-61208-279-0

An Environment for Implementing and Testing Routing Protocols in CARMNET Architecture

Adam Kaliszan
Chair of Communication and Computer Networks
Poznan University of Technology
Poznań, Poland
e-mail: adam.kaliszan@gmail.com

Mariusz Glabowski
Chair of Communication and Computer Networks
Poznan University of Technology
Poznań, Poland
e-mail: mariusz.glabowski@put.poznan.pl

***Abstract*—The paper proposes a new development environment for implementing and testing multi-criteria routing protocols in wireless mesh networks considered under the CARrier-grade delay-aware resource management for wireless multi-hop/Mesh NETworks (CARMNET) project. The paper presents the justification for choosing Optimized Link-State Routing (OLSR) protocol as the reference CARMNET routing protocol for further extension from a single-criterion to the multi-criteria protocol. The existing software router architectures (Quagga, eXtensible Open Router Platform – XORP, and BIRD) are evaluated. A software and hardware architecture of a single wireless node as well as a network testbed are described. The particular phases of the proposed implementation of the modified multi-criteria OLSR protocol in the nodes of testbed environment are presented. Two types of nodes that differ in processor architecture and energy efficiency are proposed for the testbed. The high performance nodes will be used during the development of the routing protocol, while the low performance nodes (energy efficient nodes) will be used during testing the effectiveness of the protocol elaborated. The developed architecture of the testbed network provides the ability to separate the resources exploited by control functions (routing) and data plane functions (forwarding).**

Keywords-routing protocols; mobile ad hoc network; optimized link state routing protocol; software router.

I. INTRODUCTION

Modern wireless mesh networks use multihop transmission in order to provide communication between the nodes that are not in direct transmission range. Currently, the mesh networks are used to extend the range of telecom operators' wireless networks. According to this approach, assumed also in CARMNET project [1], the nodes without direct access to the Internet can use neighboring wireless nodes' (mobile or stationary) help in order to get access to the operators' networks. However, in order to make mesh networks an important solution for telecom operators, development of cross-layer resource management framework is required. Current research on mesh networks focuses mostly on: traffic stream classification, packet scheduling, buffer memory management, routing, and mobility management. One of the key problem in multihop wireless networks, especially in networks that serve heterogeneous traffic, is the problem of optimal routing.

Within the activities related to CARMNET project, a new routing protocol that allows for multi-criteria path selection will be proposed [2]. According to the project's assumptions, the new protocol will be based on OLSR protocol. The new protocol will be implemented in testbeds located in Poznan University of Technology and Scuola Universitaria Professionale SV Italiana (SUPSI) in Lugano, in order to verify algorithmic assumption taken in the proposed protocol, and to perform complex tests of the protocol's efficiency.

The main aim of the paper is to present the proposed implementation platform for elaborating multi-criteria routing protocol. The further part of the paper is organized as follows. In Section II, the choice of OLSR protocol as the reference protocol is justified. This protocol will be modified, in order to allow for multi-criteria path selection. In Section III, the existing software router architectures are described. The decision of reusing the existing implementation of OLSRd [3] and Quagga software router architecture [4] is justified. In Section IV, the software development environment and the hardware testbed, used for implementing and testing multi-criteria routing protocol, are presented. Section V concludes the paper.

II. CHOICE OF ROUTING PROTOCOL

In CARMNET project it is assumed that there is a necessity of implementing a routing protocol, which is capable of building the routing tables including not only the best path but a set of k best paths that lead to the known IP networks/nodes in wireless mesh networks [2]. The paths in the set will be selected as the subsequent shortest paths to the specified destination, based on one of the k-shortest paths algorithms [2]. The paths are determined according to main criterion, e.g., delay, and they include additional criteria (metrics). The additional metrics will be useful in order to choose the best path, that fulfills the criteria for a given traffic stream. The criteria will correspond to the Quality of Service (QoS) requirements for all traffic classes offered in the CARMNET network. An example of the criteria can be delay, a number of hops, link reliability or link load. It is assumed in the CARMNET project that – due to a large number of already elaborated routing protocols

 ISBN: 978-1-61208-279-0

for wireless networks – the specified requirements for the routing protocol can be satisfied by appropriate modification of one of the existing routing protocol (there is no need for design of a completely new routing protocol). The following evaluation criteria were assumed during searching the protocol that can form the basis for further modification:

- The protocol has to ensure the possibility of changing the path determination algorithm,
- The protocol has to ensure the backward compatibility after extension of its functionality,
- The protocol has to ensure the possibility of additional information flooding,
- The protocol should be based on a simple finite state machine,
- The protocol should demonstrate high resistance to packet loss. The loss of a packet as well as receiving an out-of-order packet cannot force the state machine to re-initialize,
- The protocol should provide the possibility of skipping certain nodes during the process of path determination.

As the result of analysing the existing routing protocols for wireless network, the OLSR routing protocol was chosen. Only this protocol fulfills all the criteria presented above.

In OLSR protocol, the Dijkstra's algorithm is used for shortest paths tree determination. This algorithm can be replaced by one of the algorithms for set of k shortest paths determination [8]. It is possible thanks to the OLSR protocol architecture, which allows to extend algorithm functionality in an easy way. The protocol consists of the core part, which is responsible for the main functionality of the protocol, and additional modules, which enable the protocol to be extended. Additionally, this approach enables cooperation between the nodes with and without extensions. According to the OLSR specification, the nodes that do not support protocol extensions are taking part in forwarding additional information (not understandable for them) in a transparent way. In CARMNET project, the OLSR extensions will be used to flooding additional information about links' and nodes' parameters as well as to elaborate a new shortest path algorithm. The specified extensions will result in a new *multi-criteria OLSR* protocol (the flooding mechanism was used in Phosphorus project [9] for broadcasting information about grid resources in Open Shortest Path First version 2 (OSPFv2) link state advertisements). Additionally, the knowledge about links' and nodes' parameters can be used by admission control algorithms.

In the OLSR protocol specification we can observe significant state machine simplification in comparison to other link state protocols. There are only three states defined (NOT_NEIGH, SYM_NEIGH, MPR_NEIGH) [8] in the process of forming OLSR adjacencies (neighbor relationship). This simplification makes easier the implementation, testing and extending of OLSR protocol. In OLSR, the sequence numbers were used to protect the protocol against numerous packet loss and out-of-order packets delivery, occurring in wireless networks. However, in OLSR protocol, in contrast to OSPF protocol, there is no handshake process (sequence numbers synchronization) between two nodes before the routing information exchange begins [10]. In OSPF protocol, there is a special bit that specifies whether neighboring nodes are synchronized or not. After receiving a packet with a wrong value of the synchronization bit, the synchronization procedure re-starts. In the case of OLSR protocol, the last synchronization number (*Packet Sequence Number*) is remembered for 30 seconds. After exceeding this time period, the packet with any value of Packet Sequence Number is accepted. Further, the node accepts only the packets with Packet Sequence Number values greater than the value included in the last message received.

The OLSR protocol allows also to disable forwarding feature in certain of the nodes: these nodes will not forward the packets that are not addressed directly to them. This feature is very useful for the nodes with limited power resources. In order to inform about activation/deactivation of the forwarding feature in the OLSR node, the node sets an appropriate value of the *willingness* field in hello messages sent to its neighbors.

The OLSR protocol implements also an effective algorithm for distributing the routing messages among the nodes. According to this algorithm, each node determines the so-called Multi Point Relays (MPRs) among their neighbors. Subsequently, the nodes send the routing messages only to the selected MPRs, not to all neighbors. The OLSR protocol limits the number of MPR nodes, giving also a possibility of using redundant nodes in order to increase network reliability.

III. SOFTWARE ROUTER ARCHITECTURE

Nowadays, the most popular software router implementations are Quagga [4] and eXtensible Open source Routing Platform (XORP) [11], [12]. Quagga is an unofficial successor of Zebra [5]. The official successor of Zebra project is the commercial product called ZebOS [6]. Because of the General Public Licence (GPL) on which Zebra was based and nowadays Quagga is developed, the Quagga is the most popular software. Quagga architecture has been applied among others in commercial product Vyatta (since version 4.0) [7].

Both solutions (Zebra and XORP projects) have modular architecture. Each module (routing module), is responsible for all functions related to a specific routing protocol implemented in it. Additionally, in both solution, there is a special module responsible for communication between the routing modules and an operating system's Application Programming Interface (API). A new architecture of a software router is proposed within the project BIRD [13]. However, the BIRD project is in initial phase of implementation.

 ISBN: 978-1-61208-279-0

Unfortunately, there is no OLSR protocol implementation in any of the three existing software router projects. There is only one free open source code OLSR implementation, called OLSRd [3]. The OLSRd implementation allows to add new plugins including OLSR protocol extensions, in accordance with OLSR protocol philosophy. Among the implemented hitherto plugins, there is also the plugin for communication with Quagga software router.

Please note that in OLSRd implementation, the program uses directly operating system's API in order to get information about the node's (router's) interfaces and in order to add new paths to the node's routing table. Additionally, this implementation of OLSR protocol uses directly the router's interfaces to send/receive OLSR messages. Such an approach means that the routing protocol (control plane) has to be run on each node performing packets forwarding (data plane). This means, that in OLSRd implementation, the routing protocol functions cannot be transferred to a separate machine, in accordance with the concept proposed in [14].

The OLSRd implementation of the OLSR protocol has no command line interpreter. Therefore, it is impossible to change the parameters of the protocol (the router configuration) during its operation. The only method of setting the parameters of the protocol is to load them directly from a file, during the protocol start-up. Each change in the router's configuration requires modification of the configuration file, and, subsequently, restart of the protocol's software. In the configuration file, in addition to the parameters of the protocol, a list of the required plugins (extending the OLSR protocol functionality) is also included.

Possibility of creating plugins eliminates the need for modification of the core part of the OLSR protocol. Unfortunately, the existing plugins' API documentation [15] is outdated and therefore, in order to create a new plugin it is necessary to study the source code of the OLSRd software. Analysis of the entire OLSRd project's code will be required in one of four phases of the works related to both the implementation of the OLSR protocol and to its modifications.

In the first phase, the OLSRd source code will be used directly. In order to adapt the OLSR protocol to the needs of CARMNET project, a special OLSRd plugin, extending the functionality of the OLSR protocol, will be implemented. The software architecture that will be built in the first phase of protocol development is shown in Figure 1. In Figure 1, the plugin extending the OLSRd functionality is depicted as a dotted line rectangle.

The second phase of the implementation of the OLSR protocol will include its co-operation with other routing protocols. This goal will be achieved due to the application and possible revision of the plugin allowing OLSRd software to communicate directly with the Quagga software. The Quagga software router architecture is shown in Figure 2. The Quagga software consists of many modules, each of

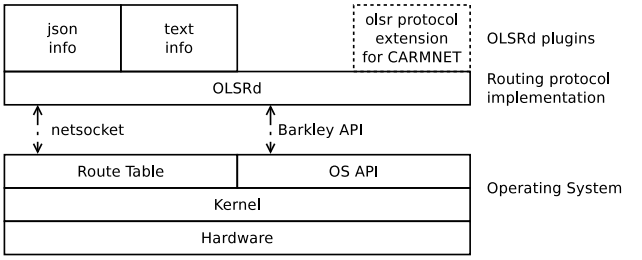


Figure 1. Extended OLSRd software architecture in the first phase of implementation

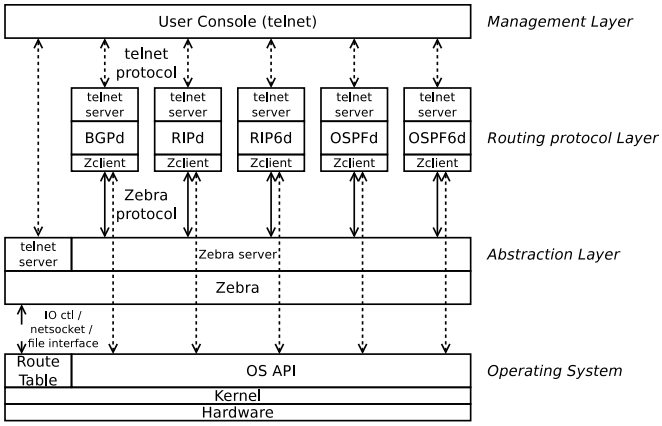


Figure 2. Quagga software router architecture

them is responsible for handling different routing protocol (BGPd daemon is responsible for Border Gateway Protocol BGP, RIPd and RIP6d daemons are responsible for Routing Information Protocol for Internet Protocol version 4 (IPv4) and Internet Protocol version 6 (IPv6), etc.). In contrast to the architecture proposed in OLSRd, the modules in Quagga do not refer directly to the operating system's API in order to retrieve information about interfaces and their addresses or in order to add new entries to the routing table. The activities listed above are executed by Zebra module that provides a bidirectional abstraction for the other modules. Communication between the modules is provided by Zebra protocol. Figure 3 shows the architecture of the system, obtained in the second phase of OLSR protocol implementation, after combining Quagga software with OLSRd software.

The software developed in the Quagga project was optimized for sake of its performance, imposing a specific approach to programming as well as the C programming language. For the typical network layer tasks (sending and receiving packets), Berkeley's socket API [16] was used in the Quagga. Consequently, also in this solution the routing protocol software (control plane) has to be running on the same machine the interfaces of which are used. The direct access of OLSRd module to the network interfaces is indicated in Figures 2 and 3 using dotted arrows.

Another approach, which will be used in phase 3 of OLSR

 ISBN: 978-1-61208-279-0

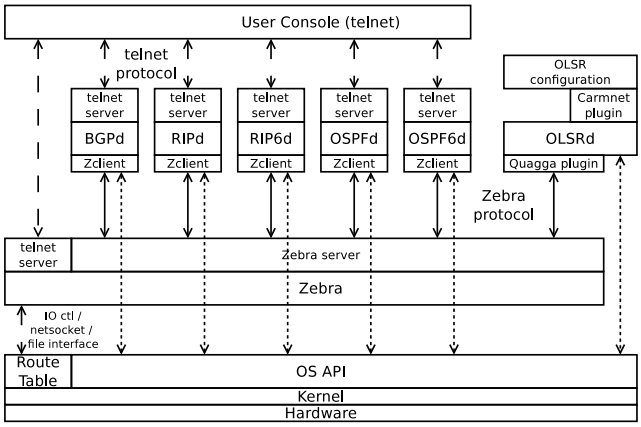


Figure 3. Software architecture in the second phase of OLSR protocol implementation

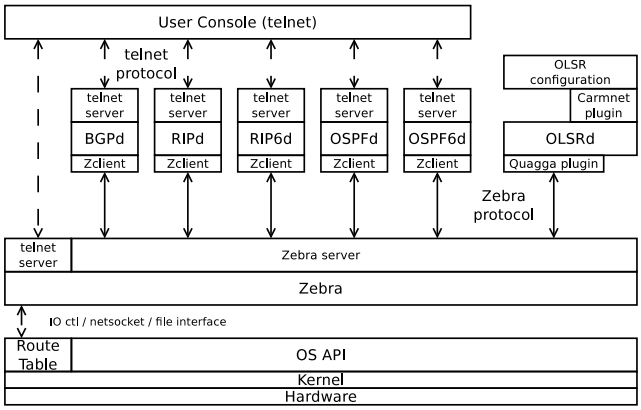


Figure 4. Software architecture in the third phase of OLSR protocol implementation: Zebra module supports sending and receiving packages for other modules

protocol's implementation and modification, was proposed in work [14]. In [14], the concept of modification of Zebra protocol and Zebra module is described. The modification is based on extending the Zebra's functionality by sending and receiving routing protocol's messages. This modification makes it possible to transfer the modules responsible for the routing from the routers to the separate machine. The transfer of the routing protocol to the separate high performance machine allows to perform many interesting experiments. Since we are not limited by (frequently) low computing power of the nodes performing data plane, it is possible to study the influence of complex routing algorithms on network performance. In addition, the transfer of the routing protocol to the separate machine does not enforce the efficient implementation of the protocol when being in experimental stage.

In order to fully integrate the routing protocol software with Quagga architecture, in phase 4 of the implementation a new module for Quagga project will be written. This module will be responsible for supporting both, the reference OLSR protocol and the proposed multi-criteria OLSR protocol. The module will use the same libraries as the other Quagga's modules and will offer a command line interpreter. The use of libraries from Quagga project will significantly reduce the size of the code. In this way it will be possible to install this software on a simple router, supported by the Open WRT system (the concept of a network consisting of devices that support OpenWRT will be presented in the next section).

IV. TESTBED NETWORK

The concept of a testbed network is shown in Figure 5. This network consists of nodes and an internal server that will facilitate the process of software development. All the nodes and the internal server are connected via a network, denoted in Figure 5 as an internal Local Area Network (LAN). This network is invisible to the routing protocols running on the nodes. The nodes are not able to use the internal LAN to exchange packets between each other.

As proposed in the previous section, the software architecture of the routing protocol will be implemented in two types of nodes: the high performance nodes, built using x86 processors, and the energy efficient nodes, built using Reduced Instruction Set Computer (RISC) processors. These nodes will then be used to build the network, enabling the testing and performance examination of protocols, implemented within the CARMNET project. The use of the nodes that were built using x86 processors (hereinafter referred to as x86 nodes) facilitates the implementation of the software, however, such nodes are more expensive and consumes more energy with respect to the nodes using RISC processors.

The main advantage of x86 nodes is that the developing software for this kind of the nodes does not require cross-compilation. The source code can be compiled directly on the developer's workstation or within the node. The simplest solution for x86 nodes is the native compilation of the software at the nodes. In the case of native compilation it is only required to provide the node with a source code. In the proposed architecture, providing software code to individual nodes will take place via an internal development server that grants an access to the repository server. An additional requirement for native compilation is the need for installing the necessary compilers and software libraries at each of the nodes in the network. Consequently, the large memory resources are required in all the nodes during the software compilation. In the case of x86 nodes, each of them is equipped with: a Solid State Drive (SSD) with the capacity of 24 GB, 4 GB of Random Access Memory (RAM) and double core ATOM Central Processing Unit (CPU). Currently, these CPUs offer the best ratio of performance to energy consumption among all x86 CPUs. In addition, the ability of installing a secondary hard disk using Serial Advanced Technology Attachment (SATA) interface in x86 nodes is provided. The interface mini Peripheral Component Interconnect (mPCIE) is used for connecting the primary

 ISBN: 978-1-61208-279-0

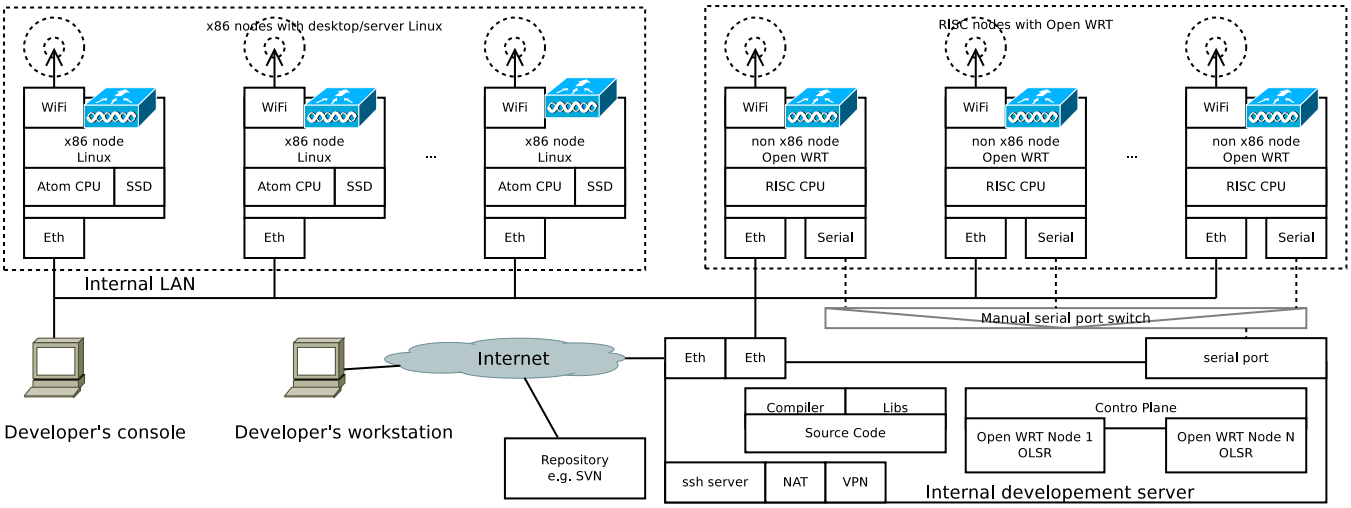


Figure 5. A testbed for development and testing of routing protocols

SSD and for connecting the Wireless Fidelity (WiFi) card. The presented node's configuration allows for installing Linux in both, "server" and "desktop" version. It is assumed that in order to speed up starting the Linux system, a desktop environment will not be installed.

The presented architecture of the nodes based on x86 processors simplifies the process of software development and testing the routing protocols. Due to the high energy requirements and the price of a single node this architecture will not be used in practical deployments. For large networks it is intended to use low cost nodes based on RISC CPUs. It was assumed that these nodes will support Open WRT [17]. Open WRT is a Linux distribution dedicated for routers. It is characterized by small hardware requirements, since a firmware image (including kernel but without WiFi support) often does not occupy more than 2 MB of memory. This is very important because most of the wireless routers available on the market are equipped with a flash memory (for storing operating system's firmware) with the capacity of 8 MB. One of the main advantages of the Open WRT is its ability to run on wide variety of processor architectures. This is possible owing to the fact that the software has an open code, which can be compiled into binary code designed for a specific platform. The compilation is done using a set of tools and libraries called buildchain. A buildchain, used in the Open WRT, is based on buildroot project [18].

In addition, the Open WRT supports ipkg package system, which allows for subsequent installation of software packages without the need for uploading the entire firmware image to a node. Despite the possibility of installing ipkg in flash or RAM memory, such a solution is often unfeasible due to the limited resources of these memories. Therefore, the Open WRT nodes will not be used for testing the routing protocol software being in the second phase of its development.

The nodes with Open WRT may be applied to test the software that is in the third phase of the development. In this phase, the control plane (responsible for routing) and the data plane (responsible for forwarding) are separated [14]. However, the transfer of the control plane from the nodes to the separate machine has some disadvantages. The nodes without the control plane are useless outside the testbed network.

The full autonomy of the nodes is achieved only in the fourth phase of the software development of OLSR routing protocol. In addition, an integration with Quagga software, completed in this phase, provides the minimum size of the module that support OLSR protocol as well as the possibility of launching it in simple routers supported by the Open WRT. The software obtained in the fourth phase will be characterized by the highest performance.

Let us notice that the process of upgrading the operating system at the nodes managed by OpenWRT may be quite complicated. In the case of certain of the nodes it may be necessary to use an additional connection to the node via a serial port. The reason for this is the application of a bootloader that works only with the serial port. The bootloader is used as the easiest way to upload a new operating system image to the node. Unfortunately, uploading the operating system via the serial port is not comfortable: the serial port offers low bit rate and, in addition, requires direct physical access to the device.

The described architecture for the implementation and testing of routing protocols foresees two working modes of the developers. In the first mode, the developers work directly on the internal server that is connected to the same LAN as the nodes (the developer's computer is just a console). The source code can be compiled on the internal development server or on each network node (see Figure 5). The machine, which is building the software, has to have

 ISBN: 978-1-61208-279-0

an access to the source code. One of the simplest yet most effective ways to provide access to the source code is to use the repository system, that is located on a server attached to the Internet. For this purpose, the internal development server performs the function of network address translation (NAT). In the case of teleworking, the access to the internal LAN is secured by a Virtual Private Networks (VPN) tunel. The mode, in which the developer has physical access to the internal LAN, has been denoted in Figure 5 by placing the computer labeled as "developer's console".

In the second mode, the developers are working without any access to the internal LAN. They only have access to the source code located on a repository server, which is connected to a public network. The compilation of the source code can be done either on the dedicated internal development server or on the programmer's workstation. The second mode is denoted in Figure 5 by the computer icon labeled as a "developer's workstation". The main limitation of this mode is that the software developers do not have any access to the network nodes, and, they are not able to perform any tests.

The final component of the proposed architecture for the implementation and testing of routing protocols are repositories that store i.a. the source code of the protocol software. According to the testbed network's architecture presented in Figure 5, the repository server is located in a separate location (accessible via public IP address). The separation of the repository and the internal development server guarantees an access to the source code even after switching the internal development server off (or in the case of its failure). Additionally, according to this solution many testbed networks may use the common repository.

V. Conclusion and Future Work

This paper proposes a testbed network architecture for implementation and testing of multi-criteria routing protocol for wireless mesh networks. The proposed testbed network architecture is programmer-friendly. It provides three ways of compiling the network nodes' source code: the native compilation on the individual network nodes, the compilation on the separate internal development server and the compilation on a developer's workstation.

Two types of network nodes are proposed for the testbed network: high-performance x86-based nodes and energy-efficient RISC nodes. The use of two types of the nodes allows to adapt the testbed network to the evolution of the routing protocol software. It is assumed that the software created in the first phase, due to its sub-optimal use of resources, will require high computing power that can be only provided by the nodes based on x86 architecture. Further optimization of the software will results in lower requirements for computing power in the nodes. Consequently, the final tests will be conducted in a large network, made up of the nodes with lower performance and greater energy efficiency (RISC architecture).

Acknowledgement

This work was supported by a grant CARMNET financed under the Polish-Swiss Research Programme by Switzerland through the Swiss Contribution to the enlarged European Union.

References

[1] "CARrier-grade delay-aware resource management for wireless multi-hop/Mesh NETworks," <retrieved: April, 2013>. [Online]. Available: http://www.carmnet.eu/

[2] J. C. Climaco, M. M. Pascoal, J. M. Craveirinha, and M. E. V. Captivo, "Internet packet routing: Application of a k-quickest path algorithm," *European Journal of Operational Research*, vol. 181, no. 3, pp. 1045–1054, September 2007

[3] "Olsrd homepage," <retrieved: April, 2013>. [Online]. Available: www.olsr.org/

[4] "Quagga homepage," <retrieved: April, 2013>. [Online]. Available: www.quagga.net/

[5] "Information about Zebra" <retrieved: April, 2013>. [Online]. Available: http://www.gnu.org/software/zebra/

[6] "ZebOS homepage," <retrieved: April, 2013>. [Online]. Available: http://www.ipinfusion.com/about

[7] "Vyatta homepage," <retrieved: April, 2013>. [Online]. Available: http://www.ipinfusion.com/about

[8] T. Clausen and P. Jacquet, "Optimized Link State Routing Protocol (OLSR)," RFC 3626 (Experimental), Internet Engineering Task Force, Oct. 2003, <retrieved: April, 2013>. [Online]. Available: http://www.ietf.org/rfc/rfc3626.txt

[9] "Phosphorus homepage," <retrieved: April, 2013>. [Online]. Available: http://www.ist-phosphorus.eu/

[10] R. Coltun, D. Ferguson, J. Moy, and A. Lindem, "OSPF for IPv6," RFC 5340 (Proposed Standard), Internet Engineering Task Force, Jul. 2008, <retrieved: April, 2013>. [Online]. Available: http://www.ietf.org/rfc/rfc5340.txt

[11] "Xorp homepage," <retrieved: April, 2013>. [Online]. Available: www.xorp.org/

[12] "Xorp architecture," <retrieved: April, 2013>. [Online]. Available: http://xorp.run.montefiore.ulg.ac.be/latex2wiki/design_overview

[13] "Bird homepage," <retrieved: April, 2013>. [Online]. Available: http://bird.network.cz/

[14] A. Kaliszan, M. Głąbowski, and S. Hanczewski, "A didactic platform for testing and developing routing protocols," in *Proceedings of The Third Advanced International Conference on Telecommunications*, Stuttgart, Germany, May 2012.

[15] A. Tonnesen, "Unik olsrd plugin implementation howto," 2004, <retrieved: April, 2013>. [Online]. Available: http://www.olsr.org/docs/olsrd-plugin-howto.html/

[16] B. Hall, *Beej's guide to network programming using internet sockets*, july 2012, <retrieved: April, 2013>. [Online]. Available: http://beej.us/guide/bgnet/

[17] "Open wrt homepage," <retrieved: April, 2013>. [Online]. Available: https://openwrt.org/

[18] "Build root homepage," <retrieved: April, 2013>. [Online]. Available: http://buildroot.uclibc.org/

 ISBN: 978-1-61208-279-0

A Low-Complexity Floor Determination Method Based on WiFi for Multi-Floor Buildings

Jian Shi and Yoan Shin
School of Electronic Engineering
Soongsil University
Seoul, Korea
{shijian, yashin}@ssu.ac.kr

***Abstract*—Floor determination has become an extremely urgent issue to resolve because many applications require accurate information on floor numbers to provide better localization services. This paper presents a Wi-Fi based, low-complexity floor determination method for multi-floor buildings. In this paper, the Multi-Wall-Floor (MWF) model is used in the simulation and the analysis. Simulation results show that the floor determination accuracy is nearly 100% if the deployment density of Wireless Access Points (WAPs) is sufficiently high on each floor. It is also shown that the proposed method provides a good estimation of floor determination even when only a few WAPs are implemented on each floor. In our scheme, detailed information on the WAP coordinates is not needed, except floor ID and Received Signal Strength (RSS) of each WAP. The novelty of the proposed method is that it can work in extreme conditions, where there are no WAPs on the floor.**

Keywords—Indoor positioning; Floor determination; WiFi; Wireless access point; Received signal strength.

I. INTRODUCTION

Determining a user's floor number in multi-floor buildings is still a difficult issue. As urbanization increases, more and more tall buildings will be built in cities. Many applications need accurate floor number information to provide better services. For example, such information could prevent school violence in multi-floor buildings by providing a faster and more efficient response to student alarm signals. Thus, floor determination has become extremely urgent to be resolved.

Outdoor positioning relies for the most part on GPS (Global Positioning System), which has an accuracy ranging from 1 meter to 10 meters. It is widely used in military applications such as surveillance and in civilian sectors such as scientific research, tracking, and navigation as well as e-commerce. However, due to attenuation and scattering, GPS is not suitable for indoors use. Hence, the requirements of indoor positioning systems differ from those of their outdoor counterparts. Many technologies have been proposed for indoor positioning in the last decade. Among these, Wi-Fi has attracted a lot of research effort because it is a mature and relatively low-cost technology. Wi-Fi hot spots are now becoming commonplace in city buildings. The utilization of these Wi-Fi hot spots' signals offers a feasible solution to floor determination [1-2,12-13].

Nowadays, numerous computing devices such as smart phones and tablet PCs are equipping Wi-Fi modules. Simultaneously, the floor number of each WAP can be known in advance. The following work is to find an effective floor determination method with RSS and ID information of scanned WAPs. To this end, we propose a Wi-Fi based, low-complexity floor determination method for multi-floor buildings. The rest of this paper is organized as follows. Section II summarizes current state-of-art on Wi-Fi based floor determination, and Section III discusses the system model and proposes our method. Simulation results are presented in Section IV to demonstrate the efficacy of the proposed method. The final section presents conclusion and future works.

II. CURRENT STATE-OF-ART OF WIFI-BASED FLOOR DETERMINATION

There are many technology options for floor determination such as time of arrival, angle of arrival, and RSS. This paper focuses on RSS-based choices because of cost and implementation consideration.

Some Wi-Fi based floor determination systems have been proposed [2-6]. Among these systems, fingerprinting-based systems play an important role [2-5]. In particular, Liu et al. [3] have demonstrated a Wi-Fi based indoor positioning system based on fingerprinting. In their research, their floor positioning experimental results showed that the floor determination was highly accurate closed to 100%. However, in common with any fingerprinting systems, the main disadvantage is that a database is usually required to train an accurate localization model. To create a fingerprinting localization model for use in multi-floor buildings, we need a sufficient number of sample points on each floor. This can be time-consuming and expensive. The indoor environment is also very complex, not only due to the presence of walls and floors, but also due to uncertain factors such as human activity and furniture or equipment rearrangements. The effects of such factors are difficult to test and measure with today's technologies. They also hamper the potential utility

 ISBN: 978-1-61208-279-0

of fingerprinting systems in the indoor environment. These systems are also unable to accommodate any changes in the Wi-Fi infrastructure and require a complete recalibration. The need to replace the database is another troublesome issue.

Alsehly et al. [2] have designed two different models for using Wi-Fi signals to determine the floor number in multi-floor buildings. One is called the "nearest floor algorithm." Essentially, it is a simplified solution of the well-known nearest neighbor classification algorithm used in fingerprinting. Although they used the system to simultaneously update records of Wireless Access Points (WAPs) in the database, the system still cannot overcome the aforementioned disadvantages. Their second model, called the "group variance algorithm" can be divided into three steps. The first step involves grouping the Wi-Fi RSS depending on the floor number. The second step entails calculating the floor parameters (range, variance, availability) for each floor. The last step involves selecting the floor number with the maximum number of points as the estimation result. Their experimental results showed that the group variance algorithm performed worse than the nearest floor algorithm. However, it was more reliable in areas such as washrooms and building edges where the received signal is weak. More importantly, the group variance algorithm does not require the creation of a database in advance. However, the disadvantage of this algorithm will be discussed later in Section IV.

The attenuations of horizontal and vertical signals are significantly different in buildings due to different materials used in the floors and the walls along with their different thicknesses. Generally, the attenuation of the floors is greater than that of the walls. This property may be exploited to estimate the floor number based on the characteristics of the received signals. Actually, floor determination will be easy if the signal attenuation of the floor is much larger up to the user's device unable to receive a signal or if the received signal is very small from the WAPs on the other floors. Thus, a critical issue that needs to be addressed in those buildings is the moderate loss of the floor penetration value.

Another disadvantage of the conventional systems is that none considers the effect of accidents, e.g., all the WAPs on one floor stopping working, but those on the other floors remaining operational as usual. The existing algorithm would be unable to determine the floor number in such a situation.

III. System Model and Proposed Floor Determination Method

A. System Model

Path loss greatly impacts the localization accuracy of algorithms based on the RSS. It results in varying degrees of loss when the radio signals propagate in different environments. Thus, choosing a suitable path loss model is very important. Research has shown that the Multi-Wall-Floor (MWF) path loss model might be the most precise model when compared to all the other models including the Motley-Keenan model for both office and commercial indoor topologies [7-11]. The following equation describes the MWF model.

$$L_{MWF} = L_0 + 10n\log(d) + \sum_{l=1}^{I}\sum_{k=1}^{K_{wi}} L_{wik} + \sum_{j=1}^{J}\sum_{k=1}^{K_{fj}} L_{fjk} \qquad (1)$$

where

L_0 : Pass loss at a distance of 1 meter

n : Power decay index

d : Distance between the transmitter and the receiver

L_{wik} : Attenuation due to wall type i and k^{th} traversed wall

L_{fjk} : Attenuation due to floor type j and k^{th} traversed floor

I : Number of wall types

J : Number of floor types

K_{wi} : Number of traversed walls of category I

K_{fj} : Number of traversed floors of category J

In addition, taking into account the influence of obstruction in indoor environment, we add a normal random variable N with zero-mean and variance of δ^2 to represent shadow noise. Then, the RSSI value from WAP can be written as:

$$L_r = L_{MWF} + N \qquad (2)$$

To make the simulation easier, we suppose that all the floors are the same type and all the walls are the same type in our model. We represent the received Wi-Fi signals of the user using the set $\Re$, which contains the number of signals. The set R_i contains the floor identity and the RSS value of the WAP.

$$\Re = [R_1, R_2, \cdots, R_i] \qquad (3)$$

$$R_i = [FloorID_i, Rss_i] \qquad (4)$$

B. Proposed Feedback Method

In an indoor environment, the thickness of each part of each floor and the thickness of the materials in the floors are nearly equal. In most cases, the thickness of each floor and the thickness of the material are also nearly equal. Based on the signal's propagation, the floor penetration loss value will fall in a certain range. In our research, we assume that random attenuation value due to the floor follows a constant value which is denoted as L_f . In practical applications, we can measure the attenuation value of each part of each floor of the building, and then average these values as the attenuation due to the floor. Based on the analysis described above, we propose a floor determination algorithm called the "feedback method."

 ISBN: 978-1-61208-279-0

The following steps describe how the proposed feedback method calculates the floor number during the analysis.

1) Pre-estimation step: We conduct a pre-estimation of the user's floor number. Generally, we choose those floor IDs that have occurred in the set $\Re$.

2) Feedback step: We suppose that the user is on the p^{th} floor which is part of the pre-estimated floor numbers. Based on this assumption, it is simple to determine how many floors the signal penetrated before the user received it. Then, we feedback the attenuation value of the floor to $\Re$. The feedback value equals to $L_f \times \left| FloorID_i - p \right|$.

3) Estimation step: We calculate the variance of all the RSS values for each pre-estimation. We then compare the variances of all the pre-estimation steps and select the floor number with the minimum variance value.

If the pre-estimation is correct, the feedback value will eliminate the influence of the floor attenuation. If not, the received RSS values in set $\Re$ will become more confused due to incorrect feedback. Thus, we selected the minimum variance value as the last estimation result. Compared to fingerprinting-based approaches, our proposed scheme does not require frequent updates of the attenuation values after accurate measurement, because the location and make-up of the floor will not usually change after the building has been built, except in exceptional cases. Compared to the group variance method, the proposed feedback method can handle the effects of accidents mentioned above. The group variance algorithm groups the RSS values according to the floor number and the last estimation results from these groups. Thus, this algorithm is unable to determine the number of the floor where the accident occurred. In contrast, our proposed scheme can handle the effects of accidents by selecting all the floor numbers as the pre-estimation.

IV. SIMULATION RESULTS

A. Simulation Environment

The simulations were performed on a multi-floor building model with eight floors. The detailed parameters and settings for the simulations are summarized in Table I.

TABLE I. PARAMETERS OF SIMULATION ENVIRONMENTS

Parameters	Values
Areas of each floor	36.5×22.7 m^2
Thickness of each floor/wall	40 cm, 30 cm
Height between floors	3 m
Attenuation due to each floor/wall	L_f = 25 dB, L_w =10 dB
MWF model	L_0 = 20 dB, n = 2.5 $N \sim (0, \delta^2)$, δ =3
RSS value	≥ –110 dBm

Note that, in order to reduce the workload, we used a simple model where the structure of each floor in the building is the same. From the commonly used free space propagation model for path loss at a distance of 1 meter and frequency of 2.4GHz (i.e., Wi-Fi frequency) we set L_0 = 20dB.

B. Analysis of Proposed Feedback Method

The estimated points were chosen at three random positions on the 1st, 5th, and 6th floor, and there were two WAPs on each floor.

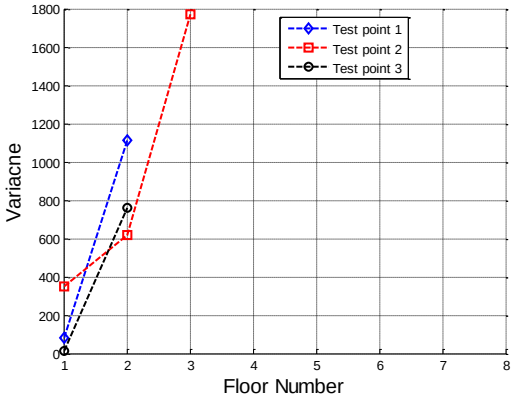


Figure 1. Simulation results for the 1st floor.

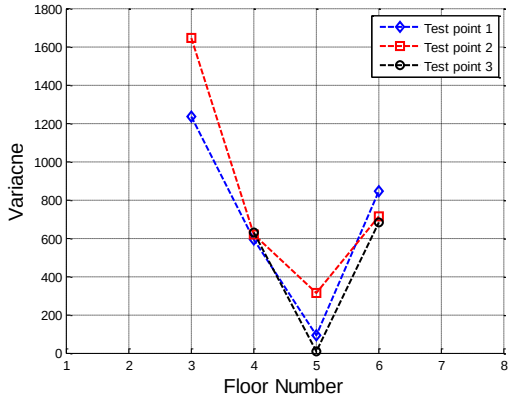


Figure 2. Simulation results for the 5th floor.

Figs. 1 and 2 show the simulation results for pre-estimation floor IDs that occur in the received Wi-Fi signals set. These figures illustrate that the lowest variance occurs when the assumed number from the pre-estimation step equals that of the actual floor number of the user.

In this paper, we also consider a specific case. In that case, all of the WAPs on the floor where an accident happens have stopped working, while WAPs on other floors can still work normally. In this case, we will choose all the floor numbers as the pre-estimation in the simulation as shown in Fig. 3.

 ISBN: 978-1-61208-279-0

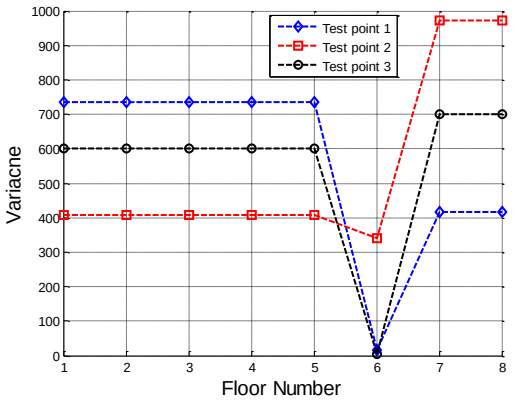


Figure 3. Simulation results for the 6th floor.

In terms of the pre-estimation, the variance maintains the same value for all the floors except the 6th floor. To explain aforementioned phenomena, we present in Table II an example with the received signals set $\Re = [R_1, R_2, R_3]$ on the 6th floor. The table clearly indicates that the variance is the same for the first five floors and the last two floors with an exception of the 6th floor.

TABLE II. EXAMPLE OF THE 6TH FLOOR

Pre-estimation	Feedback			Variance
	$R_1 = [5, Rss_1]$	$R_2 = [6, Rss_2]$	$R_3 = [7, Rss_3]$	
1	$Rss_1 + 4 \times L_f$	$Rss_2 + 5 \times L_f$	$Rss_3 + 6 \times L_f$	V_1
2	$Rss_1 + 3 \times L_f$	$Rss_2 + 4 \times L_f$	$Rss_3 + 5 \times L_f$	V_2
3	$Rss_1 + 2 \times L_f$	$Rss_2 + 3 \times L_f$	$Rss_3 + 4 \times L_f$	V_3
4	$Rss_1 + L_f$	$Rss_2 + 2 \times L_f$	$Rss_3 + 3 \times L_f$	V_4
5	Rss_1	$Rss_2 + L_f$	$Rss_3 + 2 \times L_f$	V_5
6	$Rss_1 + L_f$	Rss_2	$Rss_3 + L_f$	V_6
7	$Rss_1 + 2 \times L_f$	$Rss_2 + L_f$	Rss_3	V_7
8	$Rss_1 + 3 \times L_f$	$Rss_2 + 2 \times L_f$	$Rss_3 + L_f$	V_8

C. Comparsion with Group Variance Algorithm

As shown in Figs. 4 and 5, the comparison of the performance of the proposed feedback method with that of the conventional group variance algorithm reveals a significant change according to the number of WAPs on each floor. The simulation results are based on four different cases. Table III shows the detailed parameters.

Cases #1-4 show that the floor determination accuracy gradually deteriorates with the reduction in the number of WAPs on each floor. However, the degree of deterioration in the performance of the two methods is quite different. The proposed feedback method is always able to achieve accuracy of 95% in terms of floor determination especially for Case #1. It shows that the correct determination by the proposed method is significantly close to 100% because the deployment density of the WAPs is relatively high. In contrast, the performance of the group variance algorithm is worse than the feedback method for each case and the performance of former one decreases rather rapidly. As shown in Case #4, the group variance algorithm can realize accuracy of just 30–40%.

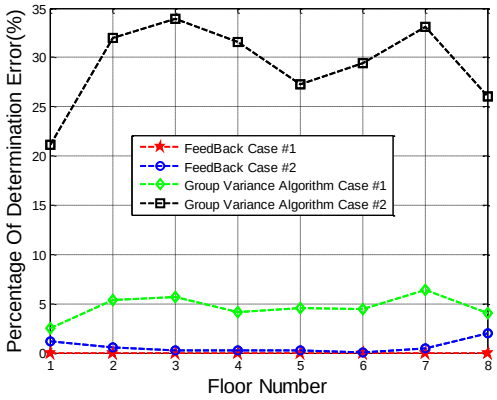


Figure 4. Comparison of the performance (Cases #1 and #2).

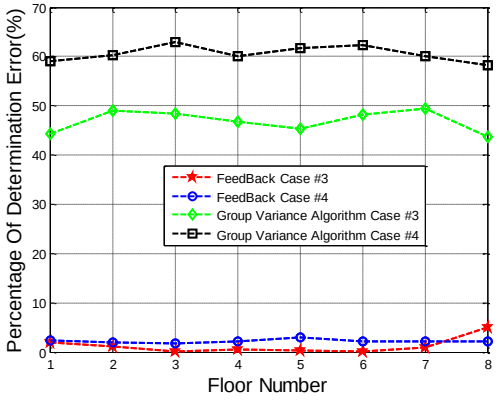


Figure 5. Comparison of the performance (Cases #3 and #4).

TABLE III. COORDINATE INFORMATION

Cases	Positions of WAPs on each floor
Case #1	(5.3 5.3) (5.3 17.4) (15.6 5.3) (15.6 17.4) (25.9 5.3) (25.9 17.4) (33.7 5.3) (33.7 17.4) (15.6 11.4)
Case #2	(10.3 10.3) (20.6 12.4) (30 5) (4 20) (30 20)
Case #3	(10.3 10.3) (20.6 12.4) (30 5)
Case #4	(10.3 10.3) (20.6 12.4)

TABLE IV. COORDINATE INFORMATION

Floor ID	Range	Variance	Availability
F1	0.5	0	0
F2	0.5	0	0
F3	0.5	0	0

The poor performance of the group variance algorithm is due to its inability to distinguish which floor the user is on when the variables shown in Table IV are present. In this

situation, the group variance algorithm cannot make a choice due to the three groups having the same points.

D. Effects of Accidents

We assumed that there are no WAPs on the 1st floor and the 6th floor, and nine WAPs on the other floors. An estimated point was chosen from 1,000 random positions on each floor.

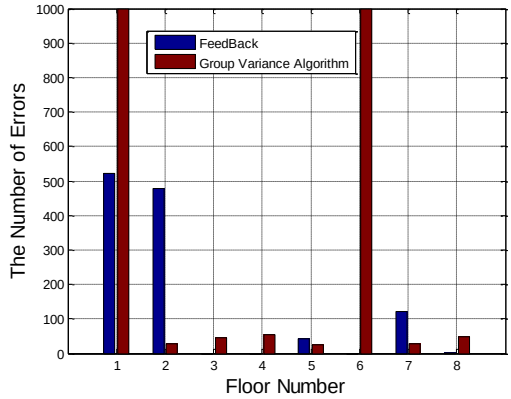


Figure 6. Comparison of the performance under extreme conditions.

Fig. 6 presents the simulation results based on this scenario. The results clearly show that the group variance algorithm was unable to adapt to this case. In contrast, the feedback method was still operational under this extreme condition. In particular, the accuracy of floor determination was high, with the change in the number of WAPs on the 6th floor seemingly causing no influence.

In order to demonstrate it clearly, we divide floor IDs into two types, as shown in Table V, since the received signals are from single direction (up-floor or down-floor), when the user locates on the accident floor which belongs to the edge building. Similarly, the signals will come from double directions (up-floor and down-floor) when the user locates on the accident floor belonging to middle building. Based on that, we can observe that the ability to deal with accidents happening on middle building is stronger than that on edge building.

TABLE V. TYPE OF FLOOR

TYPE	Floor ID
Middle building	2nd ~ 7th floor
Edge building	1st and 8th floor

Meanwhile, the accuracy of the proposed feedback method falls off sharply when the floors are close to those without WAPs (2nd, 5th and 7th floor), because the received signals are from single direction on those floors adjacent to the accident floor.

V. CONCLUSION AND FUTURE WORK

This paper presented a Wi-Fi based floor determination method for multi-floor buildings. Compared to traditional approaches, there are three advantages of the scheme proposed in this paper: robustness, simplicity, and an ability to deal with accidents. First, the simulation results showed the floor determination accuracy was nearly 100%, if the deployment density of the WAPs is sufficiently high on each floor. They also showed that the proposed method performs well in terms of floor determination, even in the presence of just two WAPs on each floor. Second, there is no need for detailed coordinate information on the WAPs in this scheme. In our research, the floor ID and the RSS value of each WAP are sufficient. Third, the final simulation results showed that this method can work under extreme conditions where there are no WAPs on the floor.

Future work will focus on perfecting the proposed algorithm for the floor determination and developing a mobile application in real environments.

ACKNOWLEDGMENT

This research was partly supported by the MSIP, Korea under the Convergence-ITRC support program (NIPA-2013-H0401-13-1004) supervised by the NIPA, and by the Human Resources Development program (No. 20114010203110) of the KETEP grant funded by the Korea Government Ministry of Knowledge Economy.

REFERENCES

[1] T. S. Perry, “Navigating the great indoors,” IEEE Spectrum, vol. 49, no. 11, Nov. 2012, pp. 15.

[2] F. Alsehly, T. Arslan, Z. Sevak, and M. King, “Indoor positioning with floor determination in multi story building,” Proc. IPIN 2011, Guimarães, Portugal, Sept. 2011, pp. 1-7, doi:10.1109/IPIN.2011.6071945.

[3] H. H. Liu and Y. N. Yang, “WiFi-based indoor positioning for multi-floor environment,” Proc. IEEE TENCON 2011, Bali, Indonesia, Nov. 2011, pp. 597-601, doi:10.1109/TENCON.2011.6129175.

[4] A. S. Al-Ahmadi, A. I. Omer, M. R. Kamarudin, and T. A. Rahman, “Multi-floor indoor positioning system using Bayesian graphical models,” Progress in Electromag. Research B, vol. 25, Sept. 2010, pp. 241-259.

[5] H. Y. Wang, V. W. Zheng, J. H. Zhao, and Q. Yang, “Indoor localization in multi-floor environments with reduced effort,” Proc. IEEE PerCom 2010, Mannheim, Germany, Mar. 2010, pp. 244-252, doi:10.1109/PERCOM.2010.5466971.

[6] S. Gansemer, U. Gropmann, and S. Hakobyan, “RSSI-based Euclidean distance algorithm for indoor positioning adapted for the use in dynamically changing WLAN environment and multi-level building,” Proc. IPIN 2010, Zurich, Switzerland, Sept. 2010, pp. 1-6, doi:10.1109/IPIN.2010.5648247.

[7] M. Lott and I. Forkel, “A multi-wall-and-floor model for indoor radio propagation,” Proc. IEEE VTC 2001-Spring, Rhodes, Greece, May 2001, pp. 464-468, doi:10.1109/VETECS.2001.944886.

[8] T. Chrysikos, G. Georgopoulos, and S. Kotsopoulos, “Wireless channel characterization for a home indoor propagation topology at 2.4 GHz,” Proc. WTS 2011, New York City, USA, Apr. 2011, pp. 1-10, doi:10.1109/WTS.2011.5960879.

[9] T. Chrysikos, G. Georgopoulos, and S. Kotsopoulos, “Impact of shadowing on wireless channel characterization for public indoor commercial topology at 2.4GHz, ” Proc. ICUMT 2010, Moscow, Russia, Oct. 2010, pp. 281-286, doi:10.1109/ICUMT.2010.5676625.

[10] J. Koo and H. Cha, “Localizing WiFi access point using signal strength, ” IEEE Commun.Lett., Feb. 2011, pp. 187-189, doi:10.1109/LCOMM.2011.121410.101379.

[11] A. Lima and L. Menezes, “Motley-Keenan model adjusted to the thickness of the wall,” Proc. SBMO/IEEE MTT-S MOC 2005,

 ISBN: 978-1-61208-279-0

Brasilia, Brazil, July 2005, pp. 180-182, doi:10.1109/IMOC.2005. 1580040.

[12] P. Brida, J. Benikovsky, and J. Machaj, "Performance Investigation of WifiLOC positioning system," Proc. TSP 2011, Budapest, Hungary, Aug. 2011, pp. 203-207, doi:10.1109/TSP.2011.6043743.

[13] J. Benikovsky, P. Brida, and J. Machaj, "Proposal of user adaptive modular localization system for ubiquitous positioning," Proc. ACIIDS 2012, Kaohsiuing, Taiwan, Mar. 2012, pp. 391-400.

 ISBN: 978-1-61208-279-0

VDTN-ToD: Routing Protocol VANET/DTN Based on Trend of Delivery

Antonio S. S. Vieira, Joao Goncalves Filho,
Joaquim Celestino Jr.
Computer Networks and Security Laboratory (LARCES)
State University of Ceara (UECE)
Fortaleza, Brazil
{sergiosviera, joao.goncalves, celestino}@larces.uece.br

Ahmed Patel
Software Technology & Management Research Center
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia (UKM)
Bangi, Sengalor, Malaysia
whinchat2010@gmail.com

***Abstract*—Providing access to the Internet or other network services to remote regions with low population density is quite complicated, since telecommunications companies may be unwilling to invest in a communications infrastructure in these locations. A possible solution to this problem is using Vehicular Ad-hoc Networks (VANET) with Delay Tolerant Networks (DTN) architecture in order to provide Internet access and other services to these regions. This is quite a challenging task because it is really difficult to predict when vehicular nodes will be in contact with each other and how long it will remain connected. In this paper , we are proposing a unique VANET/DTN routing strategy based on a new metric called Trend of Delivery (ToD). The results have shown that our proposition had better performance than other classical DTN protocols when applied to VANET. We used the Network Simulator 3 (ns-3) to implement classical DTN protocols as: Epidemic, Prophet and Spray-and-Wait in order to compare with our proposition. The simulation shows a good performance of the proposed strategy, maintaining good delivery rate while keeping overhead low, unlike purely epidemic strategy, where the overhead increases uncontrollably.**

Keywords-VANET; DTN; Routing Protocol; Fuzzy Logic; ns-3; Trend of Delivery;

I. INTRODUCTION

DTNs are a special class of networks that allow communication between regions with strong connectivity constraints, propagation delays and high error rates. In these regions, there is no way to use reliable communication and routing protocols or common standards for wireless networks as has been shown by Sadagopan et al. [1].

To overcome these constraints, the RFC4838 [2] propose that the DTN architecture can store persistent messages in a new network layer named the bundle layer until the node connectivity could be restored. This layer can provide communication between heterogeneous networks operating at different transmission media.

In the environment where we want to provide communication between the distant regions using the VANET/DTN, the Epidemic protocol [20] may achieve a high delivery rate, but its overhead is too high. The Spray-And-Wait [22] keeps overhead low, but its delivery rate may not be practical, because it controls flooding by reducing the number of copies to be made. However it does not use any smart strategy to know when it is appropriate to copy. The PROPHET [21] may have performance issues if there is no reencounters, since it is based on historical encounters.

In our paper, we propose a new VANET/DTN routing strategy, based on a new metric called the Trend of Delivery, which was created especially for this work by using fuzzy logic, in order to assist in the routing task. We also consider other recommendations proposed by Cabrera et al. [3] (the transmission range assumption, use of stale information, and an objective function in trajectory-based routing) for the development of our DTN routing protocol in order to satisfy operating constraints in a real environment. This protocol has been named the Vehicular Delay Tolerant Network - Trend of Delivery (VDTN-TD).

Bromage et al. [28] proposed a DTN routing framework based on epidemic behavior and mobility history of the nodes called the Trajectory-Assisted Routing (TAROT). In this work, the TAROT nodes only will be infected, in other words, they will receive a copy of the message, if their mobility pattern takes them closer to the destination. According to [28], when the protocol succeeds in combining intelligently replication and forwarding the messages, this leads to a reduction of overhead. Similarly to the TAROT, VDTN-TD reduces the overhead by using an intelligent forwarding strategy.

The remainder of this paper is organized as follows: Section II describes the related works. Section III covers the theoretical foundation, with a quick overview of VANET, DTN and DTN routing. The Section IV describes how ToD is calculed, also it presents VDTN-TD, the routing protocol proposed. The Section V shows the experiments and resultsobtained through simulations in ns-3. Finally, Section VI includes an outlook to future works and the conclusion.

II. RELATED WORK

DTNs are recurrent themes in the literature [3-7]. Many routing protocols have been proposed and they can be classified according to their operations [7]. Franck et al. [8] presented the attributes and requirements for VANET's operation and described how DTN can help to overcome the problems caused by high disconnection in a vehicular network. For this purpose, many tests were performed using nodes with and without vehicular DTN activated in different network conditions. The authors concluded that purely a VANET node has a high rate of packet loss and that even in a scenario with VANET/DTN packet loss is encountered due to routing loops in isolated clusters. This paper uses the combined

 ISBN: 978-1-61208-279-0

VANET/DTN to overcome disconnection problems inherent in vehicular networks.

Cheng et al. [9] proposed a DTN routing protocol called hybrid GeoDTN+Nav that explored the details of vehicular mobility and navigation systems on-board to carry the messages. According to the authors, the protocol outperforms protocols Greedy Perimeter Stateless Routing [10] and Greedy Perimeter Coordinator Routing [11] by comparing the rates of message delivery.

Besides the VDTN-ToD uses this mobility information, it also uses fuzzy logic [12] to identify mobility patterns and then decide when it is appropriate to: copy, forward or stay with the message.

III. THEORETICAL FOUNDATION

A. Vehicular Ad-hoc Networks

VANETs are a subclass of mobile ad hoc networks that emerged due to the advancement in technology for transmitting wireless networks. One of the main factors that boosted its development was the need to improve safety and traffic efficiency through communication between vehicles. Although similar to a Mobile Ad-hoc Network (MANET), a VANET has specific characteristics that the protocols, developed for MANET, are not suitable for operation in VANETs [13].

In a VANET, two categories of applications can be developed. One is focused on safety and the other should provide comfort to passengers. The first category contributes to improving the efficiency of vehicular traffic and to decrease risk situations such as sending alert messages when an accident occurs. However, the second category's focus is to make the entire trip enjoyable for the passengers. This can be achieved through applications that provide access to the Internet such as, chat with passengers of other vehicles, radios online, games, restaurants or other information relevant to the comfort of the trip [14].

B. Delay Tolerant Networks

DTN is architected by the group Delay Tolerant Network Research Group (DTNRG), a part of the Internet Research Task Force (IRTF), that enables the operation of networks operating in environments with intermittent connectivity and high delays.

In the design of the DTN architecture the following requirements have been proposed: reliable delivery, security services and flexible framework for identification of late binding. In order to attend these requirements, the architecture has included a bundle layer operating above the transport layer, where the packet of this layer is called bundle, permitting the DTN to support intercommunication among heterogeneous networks through the DTN gateways.

The DTN architecture was defined in 2007 by RFC 4838 [2] together with the specifications of the bundle protocol by RFC 5050 [15]. Later, in 2010 and 2011, they were defined in RFCs 6255 [16] and 6257 [17].

In this paper, we aim to establish communication between remote regions, in the sparse environment, using vehicles as

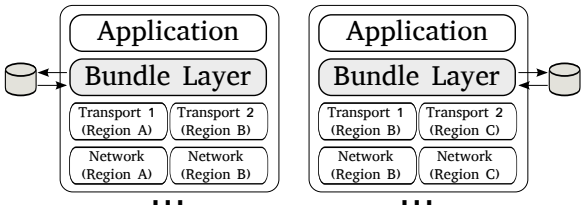


Fig. 1. Example of DTN gateways with protocol stacks specific for different regions

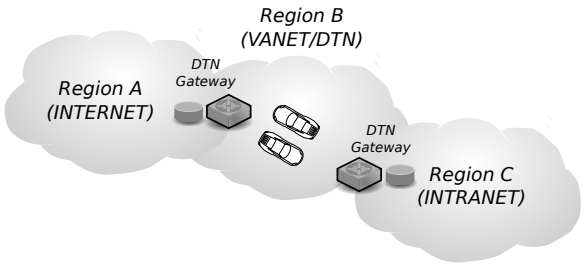


Fig. 2. Communication between remote regions

data mules [29], as can be seen in Figure 2. A region in the context of DTN, as illustrated in Figure 2, represents a communication network that comprises one or more nodes interconnected via protocols that are exclusive to their communication needs. The DTN is designed to enable communication between different regions through the DTN gateway, using the bundle layer that is able to provide such communication, through specific protocol stacks to each region as shown in Figure 1.

C. DTN Routing

A route (journey) in DTN is a sequence of communication opportunities, foreseen or unforeseen without guarantee of stability. At each contact, the message can be forwarded, copied or retained. Accordingly, the main goal of a DTN routing protocol is to increase the probability of message delivery while it aims to reduce the end-to-end delay. In challenging environments which the DTN operates, traditional routing protocols like Optimized Link State Routing Protocol (OLSR) and Ad hoc On-Demand Distance Vector Routing (AODV) do not work properly [4]. Therefore, several specific DTN protocols were proposed.

IV. ROUTING PROTOCOL VDTN-TOD

A. Trend of Delivery

Trend of Delivery is based on fuzzy logic that serves to evaluate how the mobility of nodes contributes to the delivery of a message to a fixed destination node. The ToD is composed by three linguistic variables (explained below) that are very important to the decision process, the sense ($\omega_{i,d}$), the distance ($\psi_{i,d}$) and the speed ($\tau_{i,d}$), where i denotes the node with the custody of the message and d is a border gateway of any destination region. Vehicles that possess the custody of messages must decide if the transfer will be effectuated based on these three metrics.

 ISBN: 978-1-61208-279-0

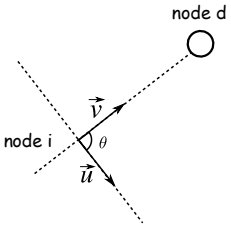


Fig. 3. Representation of the angle between the vector direction of the node and the desired direction vector

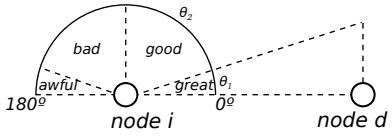


Fig. 4. Representation of categories $great$, $good$, bad and $awful$ for the variable linguistic sense

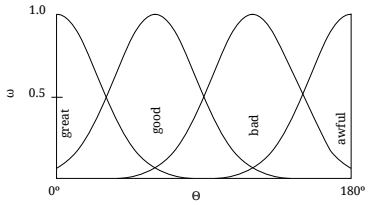


Fig. 5. Categories of variable linguistic sense

1) Sense (ω): The sense is calculated by the node that stores the message when cross with another vehicle. Its value is calculated as a function of the angle Θ formed between the direction vector $\overrightarrow{u}$ and the vector turned to the destination node $\overrightarrow{v}$, as can be seen in Figure 3. To identify the degree of importance of Θ on the approximation or departure of a node with respect to the destination, in Figure 4, we define the ranges of possible values of Θ. We propose four categories for the sense variable, based on the nebulous identification: $great$, $good$, bad and $awful$. We identify four angles associated with different categories of the sense variable in the range [0°, 180°]. We consider the same range of values for the $great$ and $awful$ classes as also to $good$ and bad classes. It is important to point out that the values of Θ_1 and Θ_2, shown in 4, are defined as a function of distance from the node i to the node d and transmission range (R). Values greater than 180° and less than or equal to 360° can be converted easily to values between 0 and 180 degrees using equation 1.

$$F(x) = 360 - x \tag{1}$$

Where x is Θ_1 or Θ_2. The ranges of the categories of ω (Figure 5) are dynamically defined as a function of the distance from node i to node d and R, this is necessary because when i reaches d the angle Θ_1 (Figure 4) gradually increases. The values of the intervals are defined according to Table I. Note that $\theta_1 + \theta_2 = 90°$.

2) Distance (ψ): For the distance linguistic variable we define four categories, namely: $very\ close$, $close$, far and $very\ far$. The pertinence functions of each variable are defined as follows:

- $very\ close$: $x < R$; $close$: $R < x \leq 2.R$;
- far: $2.R < x \leq 3.R$; $very\ far$: $x > 3.R$

Where R is the value of the transmission range and x is the distance from a node i to the destination.

3) Speed (τ): The velocity used refers to decomposed velocity for the node i ($v_{i,x}$) as the speed that indicates if the vehicle approaches or moves away from the destination. It is calculated according to equation 2.

$$v_{i,x} = v_i.cos(\theta) \tag{2}$$

For the speed linguistic variable we define three categories: low, $medium$ and $high$.

TABLE I
THE INTERVALS OF VARIABLE LINGUISTIC SENSE

	left	center	right
$\propto_{great}(x)$	$-\Theta_1$	0°	Θ_1
$\propto_{good}(x)$	Θ_1	$\frac{90°-\theta_1}{2}$	90°
$\propto_{bad}(x)$	90°	$\frac{180°-\theta_1}{2}$	$180° - \theta_1$
$\propto_{awful}(x)$	$180° - \theta_1$	180°	$180° + \theta_1$

TABLE II
FUZZY RULES FOR SETTING THE TREND OF DELIVERY

	Distance											
	Very Close			Close			Far			Very far		
	Speed			Speed			Velocity			Speed		
Sense	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low
Great	MA	MA	MA	GR	GR	VG	VG	VG	GO	GO	BA	BA
Good	MA	MA	MA	VG	VG	GO	GO	GO	BA	GO	BA	BA
Bad	MA	MA	MA	BA	BA	GO	VB	VB	BA	VB	VB	VB
Awful	MA	MA	MA	VB	BA	BA	AW	VB	VB	AW	AW	AW

These three linguistic variables $\omega_{i,d}$, $\psi_{i,d}$, $\tau_{i,d}$ are used together in order to infer the ToD using defuzzification by generating a value between 0 and 1. Furthermore, ToD is classified by seven categories: $maximum$ (MA), $great$ (GR), $very\ good$ (VG), $good$ (GO), bad (BA), $very\ bad$ (VB) and $awful$ (AW) as depicted in Table II. The ToD is considered as a maximum when the distance from a node to the destination is smaller than the transmission range.

B. Features and Operation of VDTN-ToD

The protocol VDTN-ToD uses a dissemination scheme and maintenance messages based on the location technique, Adaptive Detection Coverage, proposed by Harri et al. [18]. Each node sends a message to disseminate positioning (m_p) when travels a portion of its transmission radius. In VDTN-ToD, besides sending a message in the same way, the protocol checks in advance if it is really necessary to send a new message positioning taking into account the last positioning message sent. With the present position and speed of the vehicle. It is easy to predict the time at which a new positioning message (mp) must be sent, considering an average

 ISBN: 978-1-61208-279-0

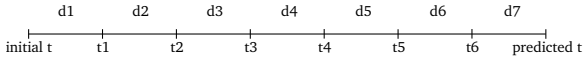


Fig. 6. Intervals of time checking to verify the need to send a new positioning message

speed of the vehicle and the transmission range. The interval between the instant of sending the initial message and the instant expected is divided into seven equal parts as shown in Figure 6.

After each interval t_i, the real position of the neighbouring vehicle (P_r) is compared to the predicted position (P_p), i.e., if the absolute difference between P_r and P_p is greater than a constant ϵ (epsilon), a new message is sent to all vehicles in the neighborhood. The m_p consists of the following fields:

- **Position-** position vector;
- **Velocity-** velocity vector;
- **Sending Time-** time sending of the message;

When a node i receives a m_p of a node j it adds a new entry in its routing table and calculates the time that both remained in contact. If during that time, node i does not receive another updated positioning message from j, node j is removed from the routing table i.

The VDTN-ToD protocol periodically checks the routing table in order to identify when a route has expired. This verification cycle is called the cycle of protocol (cp). At the beginning of each cp, the VDTN-ToD calculates the expected ToD of its neighbours for the next cycle, as well as their intended ToD, so the decision making by the protocol can be based on a configuration of the future network.

When the two nodes i and j get contact with each other and the first one has custody of a message, it compares its ToD with the one of the node j to decide which routing strategy should be used. In this sense, the strategy may be:

1) Read: $ToD(i,m)$ Trend of Delivery of the node i with the message m
2) If ($(ToD(i,m) = \{great,\ very\ good$ or $good\})$ and $(ToD(j,m) = \{great,\ very\ good$ or $good\})$ and $(ToD(i,m) \leq ToD(j,m))$), then the vehicle i copies the message m to the vehicle j, thus the VDTN-TD is intended to increase the probability of message delivery.
3) If ($(ToD(i,m) = \{bad,\ very\ bad$ or $awful\})$ and $(ToD(j,m) = \{bad,\ very\ bad$ or $awful\})$ and $(ToD(i,m) \leq ToD(j,m))$), then the vehicle i transfers the custody of message m to the vehicle j in order to delay the remoteness of the message in relation to the destination without overloading the network with unnecessary copies of the message.
4) If $(ToD(i,m)$ and $ToD(j,m)$ are opposed, for example, $good$ and bad, then
   a) If $(ToD(i,m) < ToD(j,m)$ the vehicle i transfers the custody of message m to the vehicle j.
   b) If $(ToD(i,m) > ToD(j,m)$ i remains with the message m.

When i has more than one neighbour, for each message, the one that possess the greatest ToD will be chosen among all its neighbours. This is represented mathematically as follows:

$$y = max(ToD(j,m)_{\forall j \in J, \forall m \in M}) \quad (3)$$

Therefore, the $ToD(y,m)$ will be compared with the $ToD(i,m)$ in the decision making according to the protocol operations previously seen.

The buffer management is an important aspect when using DTN protocols. We propose to use the ToD in order to manage this problem. Here, the bundle that has a higher ToD has priority over other bundles to remain in the buffer when the drop is necessary. To know the order of which bundle in the queue has to be sent, it is first checked whether any of them is a direct link, in other words, whether the neighbouring node is the destination of the bundle, if not a such case, then the bundle that has a higher ToD will be chosen.

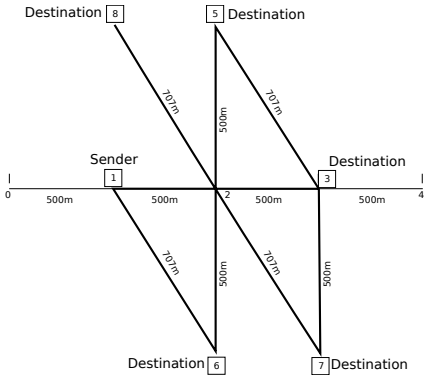


Fig. 7. Simulation scenario used in the experiments, The regions are places where are marked sender and destinations

TABLE III
ROUTING TABLE USED IN THE SIMULATION

Routes for Region 3	Routes for Regions 5, 6, 7 e 8
0-1-2-3-4	0-1-2-7
0-1-6-2-3-4	0-1-2-8
0-1-2-5-3-4	0-1-2-5
0-1-2-7-3-4	0-1-2-6
0-1-6-2-5-3-4	0-1-6-2-5

V. EXPERIMENTS AND RESULTS

A. Scenario Description

The scenario used in our experiments is shown in Figure 7. It has a fixed node (1) considered as a DTN Gateway which is responsible for sending bundles with pre-defined sizes and rates for destinations (3,5,6,7,8). This scenario was developed using the tool Simulation of Urban Mobility (SUMO) [19], where we created five routes that lead towards region 1 and five others that lead towards regions 5, 6, 7 and 8 (Table III) on the map. The thicker lines on the Figure 7 represent a road with two lanes opposed to each other where two vehicle nodes travel. The first vehicle follows the circular route 5-2-6-1-2-3-5 and the second one follows 6-2-5-3-2-1-6 throughout the simulation, other nodes randomly choose which noncircular routes of the Table III they will follow. During

 ISBN: 978-1-61208-279-0

the generation of the bundles, the DTN sender node knows the geolocation of the gateway node of the destination regions and this information is loaded into the bundle, thus the ToD calculation can be performed during the decision making of the protocol.

TABLE IV
CONFIGURATION OF EXPERIMENTS

Parameter	Configuration
Simulated Environment Area	2.000 x 1.000 m^2
Transmission Range	300 m
Maximum Speed of Nodes	40 m/s (144 km/h)
Propagation Model	Nakagami
Model Mobility	carFollowing-Krauss (SUMO Default)
Size of Bundle	256, 512, 1024 and 2048 bytes each generating respectively 196, 119, 66, 27 bundles
Simulation Time	500 seconds
Flow Vehicles	50 (generate 15 nodes), 150 (generate 29 nodes) Vehicles per Hour (VpH)
Bundle Lifetime	200 seconds
DTN Routing Protocols	Epidemic, Prophet, Binary Spray-and-Wait and VDTN-ToD
Amount of simulations for each scenario	30
Confidence Interval	95%

The experiments were performed using the Network Simulator 3.13 as described in Table IV. We implemented the following classical DTN protocols: Epidemic [20], Probabilistic Routing Protocol using History of Encounters and Transitivity (PROPHET) [21] and Spray and Wait [22] in order to compare with our VDTN-ToD.

B. Metrics

For this work we use three metrics to evaluate and compare the performance of VDTN-ToD, they are the $Delivery\ Rate$, the $Overhead$ and the $Average\ Delay$. We assume B_g for bundles generated by the source during simulation, B_c for bundles copied, B_{ct} for bundles that are transfered by custody and B_r for bundles successfully received in destination.

The Delivery Rate is calculated taking into account the equation 4 which calculates how many bundles have reached the destination.

$$delivery\ rate = \frac{B_r}{B_g} \tag{4}$$

The overhead is calculated according to equation 5, which compares the copied and transmitted bundles with the received ones.

$$overhead = \frac{(B_c + B_{ct}) - B_r}{B_g} \tag{5}$$

The average delay denotes the average time spent by the bundles in order to travel between the source and the destination regions.

C. Analysis of Results

First, we analyze the $Delivery\ Rate$. This metric is important because it reflects how the protocols behave in different situations. In a challenging network as VANET, the main problems are related to low connectivity and packet loss. In this case, it is important that a DTN routing protocol adapted to VANET use a good strategy so as to not overly consume network resources in adverse situations and behaves well especially when transferring bundles of relatively large size.

In this context, the proposed protocol achieved good results in challenging scenarios, such as shown in Figure 8. In this

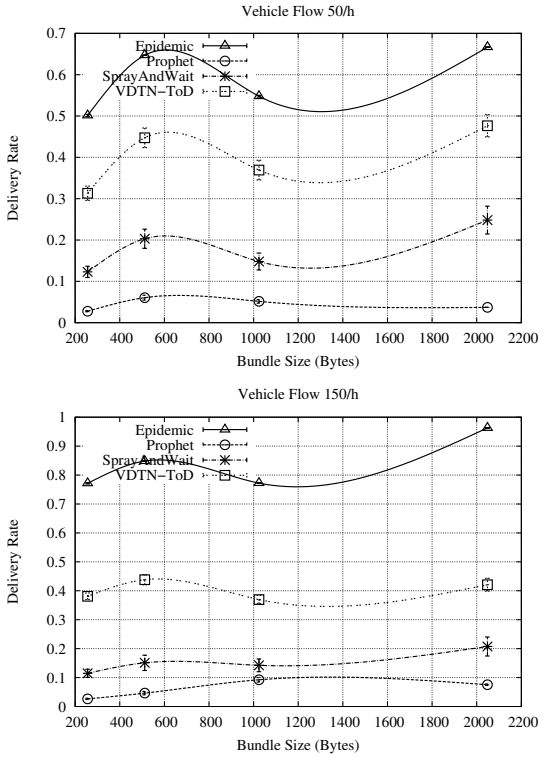


Fig. 8. Bundle Size x Delivery Rate - 50 and 150 VpH

scenario, the VDTN-ToD get a delivery rate between 30% and 50%, while the Spray-And-Wait and PROPHET stay with lower rates. In the case of the Spray-And-Wait, it do not use any mechanism to know when it should or should not spread a bundle, thus the possibility to achieve the destination is smaller than the VDTN-ToD. In the case of the PROPHET, there is a poor performance because it depended on the historical basis of the encounter, what it is difficult to build considering that in this scenario one vehicle can hardly meet another again.

We see in Figure 8 that the Epidemic protocol has the $Delivery\ Rate$ much higher than the VDTN-ToD, but as seen in Figure 9, the overhead caused by the Epidemic is extremely higher and it grows in an uncontrolled manner as the density of nodes increases, showing its low scalability. The second metric analyzed is $Overhead$. In Figure 9 we see that the overhead of VDTN-ToD is really low, becoming very close to the Spray-And-Wait (Despite its higher delivery rate). Note also that the VDTN-ToD in a more dense scenario (VpH 150) maintains the overhead close to the less dense scenario (VpH 50), unlike the Epidemic protocol that its overhead rapidly increase in a more dense scenario. We show that the VDTN-ToD can disseminate bundles efficiently to reach the destination without excessively overloading the network resources, i.e., it has a good delivery rate and is also scalable.

 ISBN: 978-1-61208-279-0

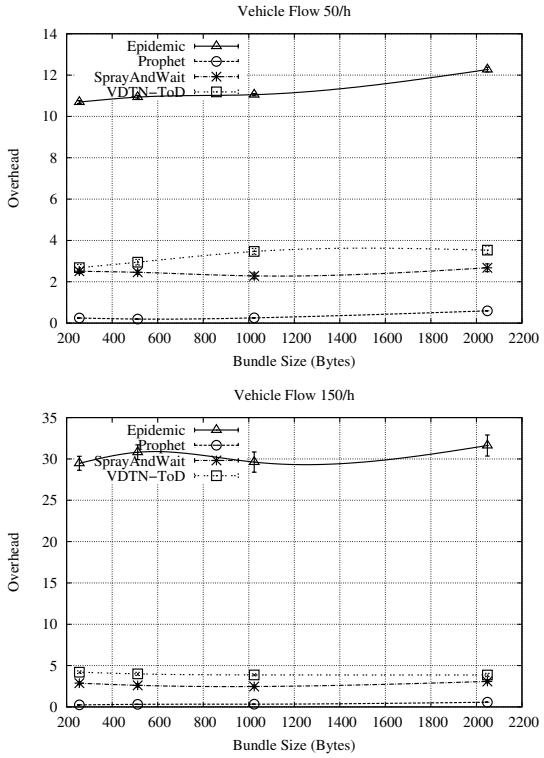


Fig. 9. Bundle Size x Overhead - 50 and 150 VpH

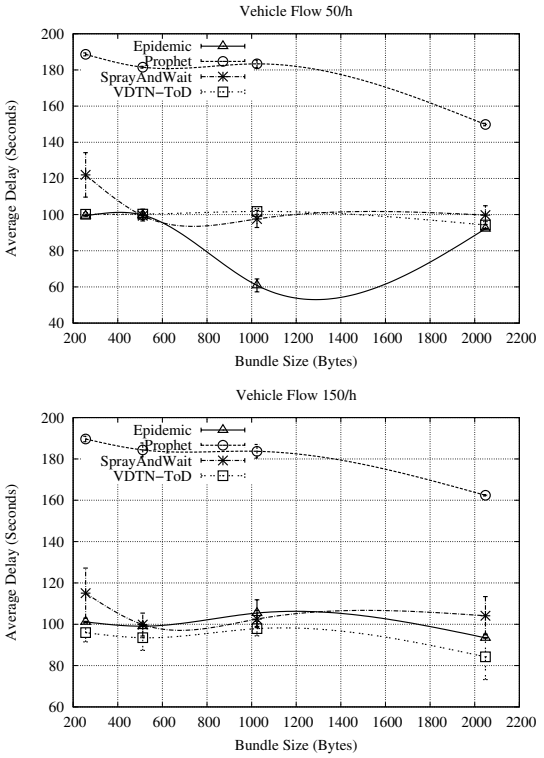


Fig. 10. Bundle Size x Average Delay - 50 and 150 VpH

The final one is the *Average Delay*. In Figure 10, in the less dense scenario (50 VpH), the Epidemic protocol achieves better results, since Epidemic node copies the bundles for all nodes that it comes in contact, the information spreads widely in this scenario, contributing to the bundle arriving quickly to the destination. When a node VDTN-ToD forward a message to a neighbouring node, the choice is aided by the proximity to the destination and speed of the neighbouring node, thereby, although a few copies, VDTN-ToD achieves better results as shown in the Figure 10 (150 VpH), where VDTN-ToD reaches a delay smaller than the Epidemic.

From our experimental results we conclude that VDTN-ToD has a good ability to adapt to the environment proposed due to its well-formulated decision making, ensuring smart forwards that cause low network overhead. This can be seen through its good delivery rate and low overhead.

VI. Conclusion and Future Works

One of the major simulation tools used to validate the DTN routing protocols is the Opportunistic Networking Environment (ONE). This simulator is used in several published studies [23] [24] [25], but there are situations where it is necessary to consider the propagation error of the messages that are inherent in wireless networks what is not provided by ONE. Differently, in this work, we use the simulator ns-3. We develop an environment capable of simulating a vehicle network together with the architecture of Delay Tolerant Networks using a module developed by Herbertsson [26]. Regarding the techniques used to create the protocol VDTN-ToD, becomes clear that we achieved great results, regarding good practices that should be employed in the development of protocols that must consider the characteristics and limitations of vehicular networks.

In general, the metric trend of delivery can be used for decision making in other DTN routing protocols. A future work is to incorporate the ToD mechanism in PROPHET and Spray-and-Wait in order to be used as a VDTN protocol. Furthermore, we will analyze the behavior of our protocol in a denser scenario, also we will evaluate perfomance when source and destination are mobile nodes. The classical DTN protocols and VDTN-ToD are available for downloading in [27].

References

[1] N. Sadagopan, F. Bai, B. Krishnamachari, and A. Helmy, “Paths: analysis of path duration statistics and their impact on reactive manet routing protocols,” in Proceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing. ACM, 2003, pp. 245-256.

[2] V. Cerf et al. “Rfc 4838, delay-tolerant networking architecture,” IRTF DTN Research Group, 2007.

[3] V. Cabrera, F. Ros, and P. Ruiz, “Simulation-based study of common issues in vanet routing protocols,” in Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th. IEEE, 2009, pp. 1-5.

[4] T. Hossmann, T. Spyropoulos, and F. Legendre, “Know thy neighbor: Towards optimal mapping of contacts to social graphs for dtn routing,” in INFOCOM, 2010 Proceedings IEEE. IEEE, 2010, pp. 1-9.

[5] T. Spyropoulos, R. Rais, T. Turletti, K. Obraczka, and A. Vasilakos, “Routing for disruption tolerant networks: taxonomy and design,” Wireless networks, vol. 16, no. 8, 2010, pp. 2349-2370.

 ISBN: 978-1-61208-279-0

[6] E. de Oliveira, C. de Albuquerque, "Nectar: a dtn routing protocol based on neighborhood contact history," in Proceedings of the 2009 ACM symposium on Applied Computing. ACM, 2009, pp. 40-46.
[7] M. Liu, Y. Yang, and Z. Qin, "A survey of routing protocols and simulations in delay-tolerant networks," Wireless Algorithms, Systems, and Applications, 2011, pp. 243-253.
[8] L. Franck, F. Gil-Castineira, "Using delay tolerant networks for car2car communications," in Industrial Electronics, 2007. ISIE 2007. IEEE International Symposium on. IEEE, 2007, pp. 2573-2578.
[9] P. Cheng et al. "Geodtn+ nav: A hybrid geographic and dtn routing with navigation assistance in urban vehicular networks," in Proc. 1st Annual Intl. Symp. Vehicular Computing Systems, Dublin, Ireland, 2008.
[10] C. Lochert, M. Mauve, H. Füßler, and H. Hartenstein, "Geographic routing in city scenarios," ACM SIGMOBILE Mobile Computing and Communications Review, vol. 9, no. 1, 2005, pp. 69-72.
[11] B. Karp, H. Kung, "Gpsr: greedy perimeter stateless routing for wireless networks," in Proceedings of the 6th annual international conference on Mobile computing and networking. ACM, 2000, pp. 243-254.
[12] S. Sandri, C. Correa, "Lógica nebulosa," Escola de redes neurais: conselho nacional de redes neurais, vol. 5, 1999, pp. 73-90.
[13] M. Taha, Y. Hasan, "VANET-DSRC Protocol for Reliable Broadcasting of Life Safety Messages," in Proc. of the IEEE International Symposium on Signal Processing and Information Technology, 2007.
[14] H. Moustafa, Y. Zhang, Vehicular Networks: Techniques, Standards, and Applications. Auerbach Publications Boston, MA, USA, 2009.
[15] K. Scott, S. Burleigh et al., "Rfc 5050, bundle protocol specifications," IRTF DTN Research Group, 2007.
[16] M. Blanchet, "Delay-tolerant networking bundle protocol iana registries," 2011.
[17] S. Farrell, H. Weiss, S. Symington, and P. Lovell, "Bundle security protocol specification," 2011.
[18] J. Härri, C. Bonnet, and F. Filali, "Kinetic mobility management applied to vehicular ad hoc network protocols," Computer Communications, vol. 31, no. 12, 2008, pp. 2907-2924.
[19] D. Krajzewicz, G. Hertkorn, C. Rossel, and P. Wagner, "SUMO (Simulation of Urban MObility); An open-source traffic simulation," in 4th Middle East Symposium on Simulation and Modelling (MESM2002), 2002, pp. 183-187.
[20] A. Vahdat, D. Becker et al., "Epidemic routing for partially connected ad hoc networks," Technical Report CS-200006, Duke University, Tech. Rep., 2000.
[21] A. Lindgren, A. Doria, and O. Schelén, "Probabilistic routing in intermittently connected networks," ACM SIGMOBILE Mobile Computing and Communications Review, vol. 7, no. 3, 2003, pp. 19-20.
[22] T. Spyropoulos, K. Psounis, and C. Raghavendra, "Spray and wait: an efficient routing scheme for intermittently connected mobile networks," in Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking. ACM, 2005, pp. 252-259.
[23] V. Le, H. Scholten, and P. Havinga, "Towards opportunistic data dissemination in mobile phone sensor networks," in ICN 2012, The Eleventh International Conference on Networks, 2012, pp. 139-146.
[24] J. Whitbeck, Y. Lopez, J. Leguay, V. Conan, and M. De Amorim, "Push-and-track: Saving infrastructure bandwidth through opportunistic forwarding," Pervasive and Mobile Computing, 2012.
[25] N. Belblidia, M. De Amorim, L. MK Costa, J. Leguay, and V. Conan, "Part-whole dissemination of large multimedia contents in opportunistic networks," Computer Communications, 2012.
[26] F. Herbertsson, "Implementation of a delay-tolerant routing protocol in the network simulator ns-3," Ph.D. dissertation, Linköping, 2010.
[27] "Dtn routing protocols in ns-3," access date: 1 May. 2013. [Online]. Available: jeri.larces.uece.br/~joaogf/vanet/ns3Dtn.zip
[28] M. K. Bromage, J. T. Koshimoto, and K. Obraczka, "Tarot: trajectory-assisted routing for intermittently connected networks," in Proceedings of the 4th ACM workshop on Challenged networks. ACM, 2009, pp. 9-18.
[29] M. Jain and R. Patra, "Implementing delay tolerant networking," 2003.

 ISBN: 978-1-61208-279-0

Tighter Effective Bandwidth Estimation for Multifractal Network Traffic

Jeferson Wilian de Godoy Stênico and Lee Luan Ling
School of Electrical and Computer Engineering
State University of Campinas - Unicamp
Emails: jeferson, lee @decom.fee.unicamp.br

***Abstract*—In literature, several studies have shown the presence of fractal nature in a wide variety of traffic and the impact of these phenomena on network performance. In this paper, we derive a new expression for effective bandwidth estimation in order to offer better resource allocation in network planning and design, especially for network traffic with multifractal characteristics. Based on a new construction approach for conservative multiplicative cascades proposed in literature and the corresponding multifractal traffic parameters, a global scaling parameter is determined and used together with the multifractal traffic model parameters for the effective bandwidth computation. The proposed approach was validated in terms of dynamically allocated bandwidths.**

Keywords - Multifractal Traffic; Global Scaling Parameter; Effective Bandwidth.

I. INTRODUCTION

The concept of effective bandwidth provides a way to characterize the resource requirements of a connection, being a useful tool for analysis and description of traffic in networks. It is considered that the effective bandwidth is the rate of transmission of information, usually with the lower limit the average rate and upper limit the peak rate of traffic, given the Quality of Service (QoS) requirements set a priori for a given traffic flow. A very good review and perspective on effective bandwidths can be found in [3].

The effective bandwidth of a source is highly sensitive to the statistical properties of the source which frequently are not known a priori. Accurate effective bandwidth estimation depends on the how faithful is the chosen traffic model. Technically the concept of effective bandwidths is much broader than a simple measure, depending on traffic models, queue disciplines and performance criteria.

Effective bandwidths required to meet the QoS requirements also depend on the traffic characteristics. The characteristics of traffic flows in current networks make their estimation no trivial and difficult using too simplified traffic models such as Markov models. Several methods of effective bandwidth estimation have been developed for broadband network traffic flow mainly based on different traffic modeling approaches. Among them the most representative ones are the following: The estimation of effective bandwidths based on self-similar traffic modeling proposed by Norros [4]; the so-called empirical effective bandwidth proposed by Tartarelli, et al. [5] without assuming any specific statistical traffic model; effective bandwidths for ATM (Automatic Teller Machine) traffic based on Markov multi-class fluid modeling proposed by Kesidis, et al. in [6]; and finally effective bandwidths based on traffic under the VVGM (Variable Variance Gaussian Multiplier) multifractal model proposed by Krishna, et al. in [7].

There are many studies showing the high variability and fast evolution of today´s internet traffic due to new applications and control protocols, i.e., modern traffic flows present variable bursts in a wide range of time scales, in contrast to the old assumptions that bursts of traffic exist only on short time scales [8, 9]. It has been shown that these incidences of multi-scales bursts affect significantly network performance [8, 9].

More realistic modeling attempts appeared, initially for characterizing the self-similarity of Internet and Ethernet traffic [10]. Although the self-similarity has provided a plausible explanation, it has failed to justify some essential local behaviors and statistical measures of real traffic flows. Therefore, the term self-similarity generally refers to those processes which are asymptotically or the second order self-similar, or monofractal [11]. In these cases, the Hurst parameter has been widely used to provide a measure of the degree of self-similarity of traffic processes.

In order to achieve even more realistic traffic modeling, taking into account multiple scaling properties as well as providing robust description of local behavior of modern network traffic, multifractal theory was adapted and used for the building of new network traffic models. Multifractal traffic modeling has enjoyed considerable success due to its theoretical robustness, versatility and generalization capability. Some well-known multifractal models designed and used for modern network traffic modeling are: VVGM [7], VSCM (Variable Scale parameter Cauchy Multiplier) [12], MWM (Multifractal Wavelet Model) [13], AWMM (Adaptive Wavelet Based Multifractal Model) [14], and mBm (multifractional Brownian motion) [15]. No doubt have those traffic models provided a more accurate description of traffic flows and contributed to the improvement in the network simulation and design tools.

The main purpose of this work is to derive and evaluate effective bandwidth for data source under a multifractal model proposed in our previous work [1, 2]. The construction of this model has been based on a new conservative multiplicative binomial cascade with its multipliers determined by a Newton Binomial equation. The major strength of this model is its high capability of capturing major multifractal properties represented by the corresponding scaling function and moment factor.

 ISBN: 978-1-61208-279-0

Therefore, this work also intends to validate this new multifractal traffic model by comparing the efficiency of the derived effective bandwidth expression with others well-established in the literature.

The paper is organized as follows. In Section II, we present a brief description of the multifractal model proposed in [1, 2]. In Section III, we show in detail the derivation of the effective bandwidth expression. In Section IV we provide a brief summary of other effective bandwidth estimation methods used for experimental investigation. Section V is dedicated for the presentation and comparison of obtained experimental results. Finally, in Section VI we conclude.

II. MULTIFRACTAL TRAFFIC MODEL

The multifractal traffic model used in this work was proposed in our previous papers [1, 2] and in this section we present this model with enough details in order to be able to understand the follow-up.

Definition 1: A stochastic process $X(t)$ is called multifractal if it has stationary increments and satisfies:

$$E(|X(t)|^q) = c(q)t^{\tau(q)+1} = c(q)t^{\tau_0(q)} \quad (1)$$

for some positive values $q \in Q$, $[0,1] \subseteq Q$, $\tau(q)$ (scaling function) and $c(q)$ (moment factor) are functions on domain Q and are independent of t. The function $\tau(q)$, also known as the partition function, is concave with $\tau(0) = -1$. [16]

A. Multiplicative Cascades

Definition 2. A multiplicative cascade is an iterative process that fragments a given set into smaller and smaller pieces according to a geometric rule and, at the same time, distributes the total mass of the given set according to another scheme.

A.1.The Proposed Binomial Multiplicative Cascade

Based on the Definition 2, the proposed multiplicative binomial cascade distribute its masses according to the Newton Binomial expression $\binom{2^N}{k}(x)^{2^N-k}(1-x)^k$, where N is a positive integer representing the stage number of the cascade and $k = 0,1,\dots,2^N - 1$. Without losing the generality, consider an initial interval $I = [0,1]$, and let x be a real-valued random variable uniformly distributed over the interval I.

At the N^{th} stage of the cascade, the first subinterval has the mass by applying the following weighting factor on the unit mass:

$$W_{\underbrace{00\dots0}_{N\ digits}} = (x)^{2^N} + (1-x)^{2^N} \quad (2)$$

while for the remaining subintervals the weighting factors are:

$$W_{b_1 b_2 \dots b_N} = \binom{2^N}{i}(x)^{2^N-i}(1-x)^i|_{i=1,\dots,2^N-1} \quad (3)$$

where $b_1 b_2 \dots b_N$ is the binary representation of decimal numeral i, also used to denote the corresponding sub-interval at the N^{th} stage of the cascade. As consequence, it is easy to see that the cascade is mass conservative in expectation.

Considering the k^{th} stage of the cascade, each subinterval of the $(k-1)^{th}$ stage is further divided into two equal length intervals. Thus, at k^{th} stage of the cascade, the mass measure of the first interval $I_k = [0, 2^{-k}]$ is equal to:

$$\mu[I_k] = \mu[0,2^{-k}] = \mu[I_{k-1}]W_{\underbrace{00\dots0}_{k\ digits}} = \mu[0,2^{-k+1}]W_{\underbrace{00\dots0}_{k\ digits}} = \mu[0,2^{-k+1}]\left[(x_{k-1})^{2^k} + (1-x_{k-1})^{2^k}\right] \quad (4)$$

For the other intervals, we have:

$$\mu[I_k] = \mu[I_{k-1}]W_{b_1 b_2 \dots b_k} = \mu[I_{k-1}]\binom{2^k}{i}(x_{k-1})^{2^k-i}(1-x_{k-1})^i|_{i=1,\dots,2^k-1} \quad (5)$$

Notice that $x_1, x_2, x_3, \dots$ are i.i.d. random variables uniformly distributed on $[0,1]$.

Let Δt_k denote the length of each subinterval at the k^{th} stage of the cascade. Thus, the mass measure of the multifractal process on the dyadic interval of length Δt_k starting at $t = 0.b_1 \dots b_k = \sum_{i=1}^{k} b_i 2_i$ calculated as:

$$\mu(\Delta t_k) = R(b_1)R(b_1 b_2)\dots R(b_{1,\dots,}b_k) \quad (6)$$

where $R(b_{1,\dots,}b_i)$ is the multiplier of the corresponding sub-interval at the stage i of the cascade. As the multipliers are independently and identically distributed (i.i.d.), it can be shown that the expectation measurement satisfies the following scaling relationship:

$$E(X(\Delta t_k)^q) = (E(R^q))^k = \Delta t_k{}^{-log_2 E(R^q)} \quad (7)$$

Therefore, the multifractal process can be characterized through its scaling function defined by $\tau(q) = -log_2 E(R^q)$.

A.2. Capture of Multifractal Characteristics

From the Definition 1, multifractal traffic modeling consists of determination of scaling function $\tau(q)$ and the moment factor $c(q)$ [17]. This can be achieved by the product of a cascade and i.i.d. positive random variables Y's. More specifically, a multifractal traffic process model can be interpreted as the product of the random peak rate of the flow Y and the measure of burstiness $\mu(\Delta t_N)$ at the modelled time scale Δt_N. The variable Y is chosen to be independent of the cascade measure $\mu(\Delta t_k)$, then the obtained series, denoted by $X(\Delta t_N)$, satisfies the following equation:

$$E\left(X(\Delta t_N)^q\right) = E(Y^q)E(\mu(\Delta t_N)^q) = E(Y^q)\Delta t_N^{\tau_0(q)} \quad (8)$$

Analyzing Equation (8) with the definition of multifractal processes Equation (1) we can show that R and Y should be related with $\tau(q)$ and c (q), respectively, as the following:

$$\begin{cases} -log_2(E(R)^q) = \tau_0(q) \\ E(Y^q) = c(q) \end{cases} \quad (9)$$

The scaling function $\tau(q)$ can be accurately modeled by assuming that R is a random variable on [0,1] with a beta distribution $\mathrm{Beta}(\alpha, \beta)$. The beta distribution is a family of continuous probability distributions defined on the interval $[0, 1]$ parameterized by two positive, typically denoted by α and β. The beta distribution can be suited to the statistical modeling of proportions in applications where values of proportions equal to 0 or 1 do not occur. Thus, the function $\tau_0(q) := \tau(q) + 1$ related to the scaling function $\tau(q)$, can be written as [1, 2]:

$$\tau_0(q) = \log_2 \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+q)}{\Gamma(\alpha)\Gamma(\alpha+\beta+q)} \quad (10)$$

where $\Gamma(.)$ denotes the Gamma function.

In [14] and [18] the authors show that the random variable Y can be considered as having a lognormal distribution defined by its two parameters m and v. Therefore the q^{th} moment of Y is explicitly given by $E(Y^q) = e^{mq+v^2q^2/2}$. Consequently the moment factor $c(q)$ for the processes is given by [1] and [2]:

$$c(q) = e^{mq+v^2q^2/2} 2^{N\left(\log_2 \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+q)}{\Gamma(\alpha)\Gamma(\alpha+\beta+q)}\right)} \quad (11)$$

Analyzing the Equations (10) and (11), one can notice that the proposed multifractal model is fully characterized by a set of four model parameters (α, β, m, v), and the mean and variance of this traffic process are related to the model parameters, respectively, as follows:

$$E[X(t)] = e^{m+v^2/2} \quad (12)$$

$$\mathrm{var}[X(t)] = e^{2m+2v^2} 2^{2N} \left(\frac{(\alpha+\beta)(\alpha+\beta+1)}{(\alpha+1)\alpha}\right)^N e^{2m+v^2} \quad (13)$$

III. Proposed Effective Bandwidth estimation

It is well known that there is another popular way to characterize a multifractal process which is through its local Hölder exponent function [19]. The Hölder exponent also can be interpreted as a generalization of a global scaling parameter of a fractal process known as The Hurst parameter. Frequently traffic flows are assumed holding only monofractal characteristics in order to make queuing analysis simpler, i.e., adopting a simplified traffic model parameter for multifractal traffic arrivals.

A self-similar process $X(t)$ with Hurst parameter H with mean zero and variance σ^2 obeys a scaling relation of the form:

$$log\{var[X^m]\} = (2H-2)log\{m\} + log\{\sigma^2\} \quad (14)$$

where m is the aggregating parameter [20]. In particular, it was shown in [2], for the proposed cascade modeling process, one can obtain the following expression:

$$log_2\{var[X^m]\} = log_2\{e^{2m+2v^2}\} + \\ + \left\{\log_2\left(\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta-1)}\right)^N\right\} + \left\{\log_2\left(\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\right)^{-\log_2 m}\right\} \quad (15)$$

In terms of multifractal model parameter, α, β, m, v.

Comparing (15) with (14), we can establish the following equality:

$$log\{m\}(2H-2) = -log\{m\}\, log_2\left(\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\right) \quad (16)$$

Therefore,

$$H_{EG} \triangleq H = 1 - \frac{1}{2} log_2\left(\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\right) \quad (17)$$

In Equation (17) we define a global parameter H_{EG} for multifractal traffic processes, similar to the Hurst parameter H in monofractals cases. More details see [2]. Thus, considering that there is a global scaling parameter for multifractal processes, given by Equation (17), next we derive an analytical expression for effective bandwidth in terms of multifractal model parameters.

Let $X(t)$ the traffic arrival process under the proposed multifractal modeling with global scale given by H_{EG}. Assuming the stage number N in the generation of the cascade is large enough and using the fBm statistical model, we can express the mean by $E[X(\delta)] = \mu\delta$ and the variance by $var[X(\delta)] = \sigma^2\delta^{2H_{EG}}$. The moment generating function of X(t) in terms of parameters θ and δ is [21]:

$$G(\theta, \delta) = \left(e^{\mu\delta\theta + \frac{\sigma^2\delta^{2H_{EG}}\theta^2}{2}}\right) \quad (18)$$

Thus the effective bandwidth may be given as:

$$e_{b_x}(\theta, \delta) = \frac{1}{\theta\delta} log G(\theta, \delta) \quad (19)$$

Substituting the relation (18) into (19), we have:

$$e_{b_x}(\theta, \delta) = \frac{1}{\theta\delta} log\left(e^{\mu\delta\theta + \frac{\sigma^2\delta^{2H_{EG}}\theta^2}{2}}\right) \quad (20)$$

 ISBN: 978-1-61208-279-0

Thus,

$$e_{b_x}(\theta,\delta) = \frac{1}{\theta\delta}\mu\delta\theta + \frac{\sigma^2\delta^{2H_{EG}}\theta^2}{2} = \mu + \frac{\theta\sigma^2}{2}\delta^{2H_{EG}-1} \quad (21)$$

Therefore

$$e_{b_x}(\theta,\delta) = \mu + \frac{\theta\sigma^2}{2}\delta^{\left(2-log_2\left(\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\right)\right)-1} \quad (22)$$

where θ represents the asymptotic exponential decay rate of the distribution of the queue size and δ the time scale.

IV. OTHER METHODS OF EFFECTIVE BANDWIDTH

A. Norros Effective Bandwidth

Norros et al. [4] proposed an expression for effective bandwidth estimation by considering traffic with the fBm self-similar characteristics, that is:

$$\alpha = m + K(H)\sqrt{-2ln(P_{loss})}^{1/H} a^{\frac{1}{2H}} b^{-(1-H)/H} m^{\frac{1}{2H}} \quad (23)$$

where m represents the average rate of the traffic flow in (bit/s), $K(H) = H^H(1-H)^{1-H}$, a the coefficient of variance, P_{loss} the overflow probability of buffer, H Hurst Parameter and b the buffer size.

The effective bandwidth suggested by (23) takes into account the self-similarity property of traffic through its Hurst parameter, which is an appropriate alternative for most traffic flows holding long-range dependence (LRD) characteristics. The effective bandwidth estimates become much tighter when buffer size become large. For more details, see [4].

B. Empírical Effective Bandwidth

The effective bandwidth estimation by Equation (24) Proposed in [5], known as Empirical Effective Bandwidth, does not assumes any specific traffic flow model.

$$\alpha(s,t,N) = \frac{1}{st} log\hat{E}_{N_t}\left[e^{sX(0,t)}\right] \quad 0 < s; 0 < t < N_t \quad (24)$$

where $X(0,t)$ indicates the aggregated amount of arrived traffic data within a time interval t and $\hat{E}_{N_t}\left[e^{sX(0,t)}\right]$ is the data-measured moment generating function from the traffic trace with N_t samples. For both Poisson and On-Off processes, the empirical effective bandwidths are very close to their respective analytical effective bandwidths. For more details, see [5].

C. Kesidis Effective Bandwidth

In [6] Kesidis et.al. derived an expression of effective bandwidth for fluids Markov multi-class and other types of source models under ATM traffic. The authors showed that when traffic sources share a buffer system with deterministic service rate, a constraint on the tail of the buffer occupancy distribution is a linear constraint on the number of sources, i.e., for a small loss probability one can assume that each source transmits at a fixed rate called effective bandwidth.

Let m be the average rate of traffic, $s = ln(P_{loss}/B)$, P_{loss} the overflow probability of buffer and B the buffer size. The Effective Bandwidth (EB) is given by:

$$EB = m\frac{e^s-1}{s} \quad (25)$$

For more details, see [6].

D. Krishna Effective Bandwidth

Krishna et al. [7] propose an expression for calculating effective bandwidth based on the multifractal VVGM model. Also, they assumed that traffic can be characterized as fBm processes. The effective bandwidth (EB), given by (26), is written in function of the parameters θ (asymptotic exponential decay rate of the distribution of the queue size) and δ (time scale), traffic average rate m and variance σ^2, and the global scaling exponent of the VVGM model, H_{eff}.

$$EB = m + \frac{\theta\sigma^2}{2}\delta^{(2H_{eff}-1)} \quad (26)$$

For more details, see [7]

V. EXPERIEMENTAL EVALUATION

In this section, we evaluate the efficiency of the proposed effective bandwidth estimation method. Instead of obtaining a unique static bandwidth estimate for the entire traffic trace, dynamically effective bandwidth is estimated instantaneously using only traffic samples inside a sliding time window and used as the current server transmission rate.

Three real traffic traces were used in our simulation: a TCP / IP traffic trace called "lbl_tcp_3" [22], a video traffic flow called "The Simpsons" [23] (high quality video) and a traffic trace collected in a wireless network collected during the ACM SIGCOMM08 conference [24], namely "Sigcomm08". The traffic samples were aggregated under a time scale on which all three traffic flows exhibit multifractal characteristics [25]. Service is conservative, i.e., server will never remain idle if there is one or more jobs in the service node.

Table I shows some statistical information (means, variances and number of samples) of three traffic traces.

For performance comparison purposes, we also evaluate the queue system using four effective and width estimation approaches described in the previous section: the effective bandwidth proposed by Norros [4], the empirical effective bandwidth proposed by S. Tartarelli et al. [5], the effective bandwidth proposed by Kesidis [6], and the effective bandwidth proposed by Krishna [7].

Table II shows the global scaling parameter values obtained under the proposed modeling method and compares with the Hurst parameters estimated through a Whittle Estimator [26] for all three mentioned traffic traces (namely lbl_tcp_3, The Simpsons and Sigcomm08). It can be seen that numerically two global scaling parameters are close. As a result, the global scaling parameter H_{EG} can be viewed as an alternative measure for self-similarity.

 ISBN: 978-1-61208-279-0

TABLE I. MEAN, VARIANCE, SAMPLES

Traffic Trace	Mean	Variance	Samples
lbl_tcp_3	136.3555	5.7062×10^4	1.789.995
The Simpsons	6.5137×10^3	7.2420×10^6	30.334
Sigcomm08	451.9165	2.3723×10^5	1.358.782

TABLE II. HURST AND GLOBAL PARAMETER

Traffic Trace	Hurst Paramenter (H) Whittle Estimator	Global Scaling Parameter H_{EG}
lbl_tcp_3	0.8420	0.8691
The Simpsons	0.7130	0.7262
Sigcomm08	0.7650	0.7567

Figure 1 shows the necessary effective bandwidth values obtained using Equation (22) and also those using other cited methods for the lbl_tcp_3 traffic trace, 10^{-6} loss probability system performance, 64Kbytes buffer size and a sliding time window of 500 traffic samples. Notice that the proposed method outperforms other approaches requiring the lowest service rate. However, the improvement is relatively small with respect to the methods proposed in [5] and [7], and considerably remarkable in comparison with those by Norros [4] and Kesidis et.al. [6].

It is noteworthy that the method proposed by Kesidis et.al. [6] is based on Markovian traffic modeling, and it is a well-known fact that Markovian Modeling cannot fully represent traffic with multifractal characteristics [11]. As a result, the Markovian based effective bandwidth estimates may be too conservative.

Figure 2 shows the performance curves for the video traffic trace (The Simpsons), considering 10^{-6} loss probability system performance, 32Kbytes buffer size and a sliding time window of 100 traffic samples. Again, the proposed approach shows considerably better performances.

Figure 3 shows the performance curves for a wireless traffic trace (Sigcomm08), considering 10^{-6} loss probability system performance, 64Kbytes buffer size and a sliding time window of 500 traffic samples. Similar results are also observed and, once again, the proposed method prevails.

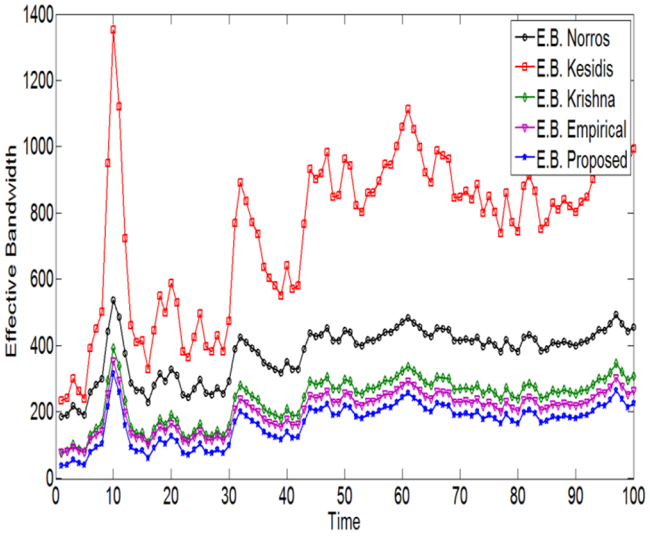


Fig.1. Effective Bandwidth for Traffic Trace lbl_tcp_3.

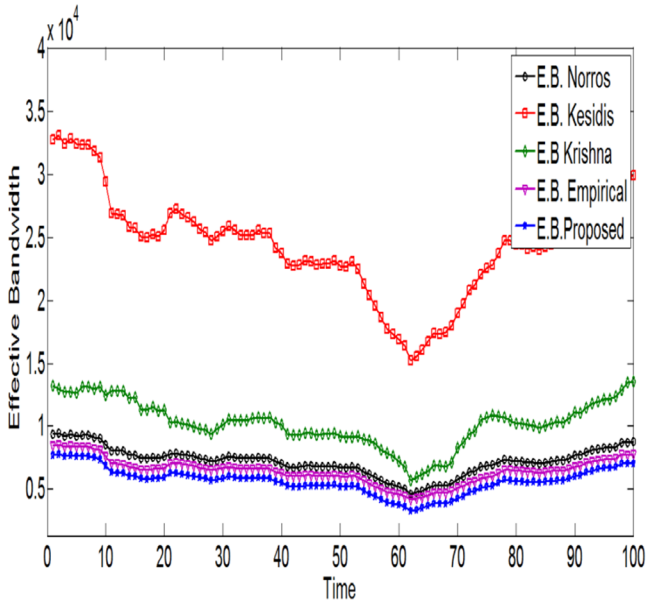


Fig.2. Effective Bandwidth for Traffic Trace Video

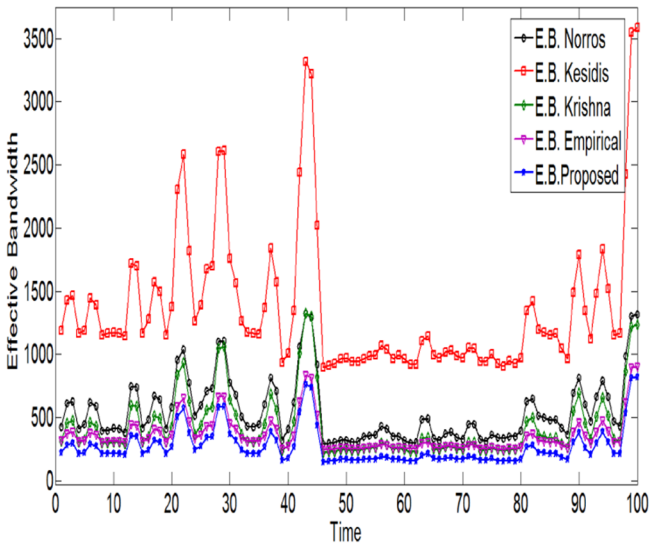


Fig.3. Effective Bandwidth for Traffic Trace Sigcomm08.

VI. CONCLUSION AND FUTURE WORK

In this work, we derived a global scaling parameter based on the multifractal traffic model presented in our recent previous work. In addition, by using this global scaling coefficient we derived an analytical expression of effective bandwidth, which take into account traffic's fractal behavior e characteristics. Experimental investigation results validated our approach showing its outstanding performance in terms of network resource usage. We also believe this global scaling parameter can be used alternatively as a measure of traffic's self-similarity.

For future work, we may investigate how efficient this global scaling parameter is in comparison with the Hurst parameter. The testing results encourage us to pursue further

 ISBN: 978-1-61208-279-0

investigation on our derived effective bandwidth expression in terms of its susceptibility and robustness with respect to the variation of traffic modeling and queue system parameters. Based on this new multifractal traffic model, as well as the experience from effective bandwidth investigation acquired from this work, new schemes for network resource allocation and admission control, possibly in real time, will be also our future research issues.

REFERENCES

[1] J. W. G. Stenico and L. L. Lee, "Modelagem de Processos Multifractais Baseada em uma Nova Cascata Conservativa Multiplicativa," XXIX Simpósio Brasileiro de Telecomunicações - SBRT 11,Vol. 1, Outubro 2011, pp.1-6, Curitiba, PR, Brasil,.

[2] J. W. G. Stenico and L. L. Lee, "A New Binomial Conservative Multiplicative Cascade Approach for Network Traffic Modeling," In 27th IEEE International Conference on Advanced Information Networking and Applications – IEEE AINA 2013, March 2013, pp. 794-801 Barcelona, Spain.

[3] F. P. Kelly, "Notes on Effective Bandwidths," In Stochastic Networks. Royal Statistical Society Lecture Notes Series, 4. 1996, pp. 141-168, Oxford University Press.

[4] I. Norros, "On The Use of Fractional Brownian Motion in The Theory of Connectionless Networks," IEEE Journal on Selected Areas in Communications, 13(6), August 1995, pp. 953-962.

[5] S. Tartarelli, M. Falkner, M. Devetsikiotis, I. Lambadaris, and S. Giordano, "Empirical Effective Bandwidths," In Proc. Of IEEE Globecom'00. Vol. 1, November/December 2000, pp. 672-678.

[6] G. Kesidis, J. Walrand, and S. Chang, "Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources," In: IEEE/ACM Transactions on Networking, vol. 1, No.4, August 1993, pp.424-428.

[7] P. M. Krishna, V. M. Gadre, and U.B. Desai, "Multifractal Based Network Traffic Modeling," Kluwer Academic Publishers, Boston, MA, December 2003.

[8] X. Jin and G. Min, "Modelling and Analysis of Priority Queueing Systems with Multi-Class Self-Similar Network Traffic: A Novel and Efficient Queue-Decomposition Approach". In IEEE Transactions on Communications, Vol.57, May 2009, pp 1444–1452.

[9] C. Parka, F. C. Hernández, L. Le, J. S. Marron, J. Park, V. Pipiras, F. D. Smith, R. L. Smith, M. Trovero, and Z. Zhu, "Long-Range Dependence Analysis of Internet Traffic". Journal of Applied Statistics. Vol 38(7), September 2011, pp. 1407-1433.

[10] J. Shunfu and Y. Wuyi, "Performance Evaluation of Self-Similar Traffic in Multimedia Wireless Communication Networks with Power Saving Class Type III in IEEE 802.16e," In IEEE International Conference on Wireless Communications, Networking and Information Security (WCNIS), November 2010, pp. 436 – 440.

[11] K. Park and W. Willinger, "Self-Similar Network Traffic and Performance Evaluation," John Wiley and Sons, New York, 2000.

[12] Z. Xu, L. Wang, and K. Wang, "A New Multifractal Model Based on Multiplicative Cascade," Information Technology Journal, Vol. 10 November 2011, pp. 452-456.

[13] R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G Baraniuk, "A Multifractal Wavelet Model with Application to Network Traffic," IEEE Transactions on Information Theory. (Special Issue on Multiscale Signal Analysis and Modeling), vol. 45, April 1999, pp. 992-1018.

[14] F. H. T. Vieira and L. L. Lee, "Adaptive Wavelet Based Multifractal Model Applied to the Effective Bandwidth Estimation of Network Traffic Flows," IET Communications, Vol. 3(6), June 2009, pp. 906 - 919.

[15] R. Peltier and J. L. Véhel, "Multifractional Brownian motion: Definition and Preliminary Results," Technical Report 2695, INRIA, August 1995, pp. 1 - 41.

[16] A. J. Fisher, L. E. Calvet, and B. B. Mandelbrot, "Multifractality of Deutschemark / US Dollar Exchange Rates" (September 15, 1997). Cowles Foundation Discussion Paper No. 1166; Sauder School of Business Working Paper

[17] I. W. C. Lee and A. O. Fapojuwo, "Stochastic Processes for Computer Network Traffic Modeling," Computer Communication, Vol. 29(1), December 2005, pp. 1 - 23.

[18] T. D. Dang, S. Molnár, and I. Maricza, "Queuing Performance Estimation for General Multifractal Traffic," International Journal of Communication Systems, vol 16 no 2, February 2003, pp. 117–136.

[19] S. Seuret and J. Lévy-Véhel, "The Local Holder Function of a Continuous Function," Applied and Computational Harmonic Analysis, Vol. 13 (3) , April 2000, pp. 263-276.

[20] J. Beran, "Long-Range Dependence" Wiley Interdisciplinary Reviews: Computational Statistics, Vol. 2 (1), January/February 2010, pp. 26-35.

[21] N. G. Duffield, "Application of Large Deviation to Performance Analysis with Long-Range Dependent Traffic," Workshop on Stochastic Modeling and Analysis of Communication Networks, page Slides of Presentation, 1998.

[22] http://ita.ee.lbl.gov/html/traces.html (Retrieved: April, 2013).

[23] http://www.cs.columbia.edu/~hgs/internet/traces.html (Retrieved: April, 2013).

[24] A. Schulman, D. Levin, and N. Spring, "CRAWDAD Data Set UMD/Sigcomm2008 (v.2009-03-02)," Downloaded from http://crawdad.cs.dartmouth.edu/umd/sigcomm2008. (Retrieved: April, 2013).

[25] R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk, "A Multifractal Wavelet Model with Application to Network Traffic". IEEE Transactions on Information Theory, Vol. 45 (3) , April 1999, pp. 992 - 1018.

[26] T. Karagiannis and M. Faloutsos, "SELFIS: A Tool For Self-Similarity and Long-Range Dependence Analysis," 1st Workshop on Fractals and Self-Similarity in Data Mining: Issues and Approaches (in KDD) Edmonton, Canada, 2002.

 ISBN: 978-1-61208-279-0

Bandwidth Reservation in the Erlang Multirate Loss Model for Elastic and Adaptive Traffic

Ioannis D. Moscholios[1], Vassilios G. Vassilakis[2], Michael D. Logothetis[3] and John S. Vardakas[4]
1. Dept. of Telecommunications Science and Technology, University of Peloponnese, Tripolis, Greece.
Email: idm@uop.gr
2. School of Computer Science & Electronic Engineering, University of Essex, Colchester, U.K.
Email: vasilak@essex.ac.uk
3. WCL, Dept. of Electrical and Computer Engineering, University of Patras, Patras, Greece.
Email: m-logo@wcl.ee.upatras.gr
4. Iquadrat, Barcelona, Spain
Email : jvardakas@iquadrat.com

***Abstract*—In this paper, we consider a single-link multirate loss system, which accommodates *K* service-classes with different traffic and peak-bandwidth requirements. Calls of each service-class arrive in the system according to a Poisson process and have an exponentially distributed service time. The *K* different service-classes are distinguished in K_e elastic service-classes and K_a adaptive service-classes ($K=K_e + K_a$). Elastic calls can compress their peak-bandwidth by simultaneously increasing their service time, while, adaptive calls can tolerate bandwidth compression without affecting their service time. The system incorporates the bandwidth reservation (BR) policy whereby we can achieve certain quality of service (QoS) for each service-class through a proper bandwidth allocation, defined by the BR parameters. To calculate, in an approximate but efficient way, Call Blocking Probabilities (CBP) and link utilization, we propose a recurrent formula for the determination of the link occupancy distribution. The accuracy of the proposed formula is verified by simulation and is found to be very satisfactory. We also show the consistency and the necessity of the new model.**

Keywords - loss system; blocking probability; reservation; elastic-adaptive traffic; recursive formula.

I. INTRODUCTION

Call-level multirate loss models of a single link that supports elastic and adaptive service-classes have received attention over the last years due to the increase of elastic and adaptive traffic in communication networks and the consequent need for QoS network assessment (e.g., [1]-[11]). Elastic traffic is composed of calls that expand their service time when their bandwidth is compressed. On the other hand, adaptive traffic refers to calls that do not alter their service time in the case of bandwidth compression. Examples of elastic traffic are generally TCP-based applications (FTP, HTTP, STMP), while examples of adaptive traffic are mostly real-time applications, like audio and video streaming, which can be transmitted with an acceptable QoS after bandwidth compression.

The analysis of a single link multirate loss system, which accommodates elastic and adaptive calls that arrive to the system according to a Poisson process and have an exponentially distributed service time has been proposed in [2]. In [2], an arriving call is accepted in the system with its peak-bandwidth requirement, when the occupied link bandwidth does not exceed the capacity of the link. Otherwise, the system accepts this new call by compressing its peak-bandwidth, together with the bandwidth of all in-service calls of all service-classes. Call blocking occurs when the maximum possible bandwidth compression is still not enough to ensure the acceptance of a new call in the system. When an in-service call, whose bandwidth is compressed, departs from the system, then the remaining in-service calls (of all service-classes) expand their bandwidth. The existence of the bandwidth compression/expansion mechanism destroys the Markov chain reversibility and therefore the model of [2] does not have a Product Form Solution (PFS). However, the calculation of CBP and link utilization is based on an approximate but recursive formula for the determination of link occupancy distribution. This formula resembles the classical Kaufman-Roberts formula used in the Erlang Multirate Loss Model (EMLM), where Poisson arriving calls have fixed bandwidth requirements (stream traffic, no bandwidth compression is permitted) and compete for the available link bandwidth under the complete sharing policy [12], [13]. To this end, we name herein the model of [2], Extended EMLM (E-EMLM). The co-existence of stream and elastic traffic in the same link has been considered in [11], but the proposed formulas are not recursive (i.e., do not resemble the Kaufman-Roberts formula).

In this paper, we incorporate into the E-EMLM the BR policy (E-EMLM/BR). The BR policy can achieve CBP equalization among service-classes (either elastic or adaptive), or guarantee a certain QoS for each service-class, by a proper selection of the BR parameters so that each service-class meets a certain bandwidth capacity. The consideration of the BR policy is of paramount importance in multirate communication networks, given that the absence of the BR policy leads to an unfair service (the less required bandwidth, the better CBP). The proposed model does not have a PFS and therefore we provide an approximate recursive formula for the determination of the link occupancy distribution. The latter not only simplifies the calculation of CBP and link utilization but also provides

 ISBN: 978-1-61208-279-0

quite satisfactory results compared to simulation. Note that a multirate loss system that accommodates Poisson arriving calls of elastic traffic only has been considered in [1], while the application of the BR policy in that model has been proposed in [14].

The remainder of this paper is as follows: In Section II, we review the E-EMLM. In Section III, we propose the E-EMLM/BR. Section IV is the evaluation section. We present analytical and simulation results of CBP and link utilization of the E-EMLM/BR. We also provide analytical results of the E-EMLM for comparison. We conclude in Section V.

II. REVIEW OF THE E-EMLM

A. Review of the E-EMLM

Consider a single link of capacity C bandwidth units (b.u.) that accommodates calls of K service-classes. Let K_e and K_a be the set of elastic and adaptive service-classes ($K_e + K_a = K$), respectively. A call of service-class k (k =1,…,K) follows a Poisson process with arrival rate λ_k and has a peak-bandwidth requirement of b_k b.u. (integer value). Let j be the occupied link bandwidth when a new service-class k call arrives in the link. If $j + b_k \le C$, the call is accepted in the system with its b_k b.u. and remains in the system for an exponentially distributed service time with mean μ_k^{-1}. The new service-class k call is blocked and lost if $j + b_k > T$, where T is the limit (in b.u.) up to which bandwidth compression is permitted. If $T \ge j + b_k > C$ the new call is accepted in the system. However, the assigned bandwidth of all in-service calls, together with the peak-bandwidth requirement of the new call is compressed. After the bandwidth compression of all calls (new and in-service) the system state becomes $j = C$. The compressed bandwidth of the new service-class k call is calculated by:

$$b_k^{'} = r b_k = \frac{C}{j^{'}} b_k \tag{1}$$

where $r \equiv r(\boldsymbol{n}) = C/j^{'}$, $j^{'} = j + b_k = \boldsymbol{nb} + b_k$, $\boldsymbol{n}=(n_1, n_2,\ldots,n_k,\ldots, n_K)$, n_k is the number of in-service calls of service-class k, $\boldsymbol{b}=(b_1, b_2, \ldots,b_k,\ldots, b_K)$ and $j = \sum_{k=1}^{K} n_k b_k = \boldsymbol{nb}$.

Similarly, the compressed bandwidth of all in-service calls is equal to $b_i^{'} = \frac{C}{j^{'}} b_i$ for i=1,…,K. The minimum bandwidth of a service-class k call is given by:

$$b_{k,\min}^{'} = r_{\min} b_k = \frac{C}{T} b_k \tag{2}$$

After the bandwidth compression, all elastic calls increase their service time so that the product (service time) by (bandwidth) remains constant.

The mechanism of bandwidth compression/expansion destroys reversibility in the E-EMLM and therefore no PFS exists. However, in [2] an approximate recursive formula is proposed which determines the link occupancy distribution, $G(j)$, (unnormalized values):

$$G(j) = \begin{cases} 1 & \text{for } j = 0 \\ \frac{1}{\min(j,C)} \sum_{k \in K_e} \alpha_k b_k G(j - b_k) + & \\ \frac{1}{j} \sum_{k \in K_a} \alpha_k b_k G(j - b_k) & \text{for } j = 1,\ldots,T \\ 0 & \text{for } j < 0 \end{cases} \tag{3}$$

where $\alpha_k = \lambda_k/\mu_k$ is the offered traffic-load (in erl) of service-class k.

Based on (3) we can calculate CBP and the link utilization, as follows:

1) The CBP of service-class k, B_k:

$$B_k = \sum_{j=T-b_k+1}^{T} G^{-1} G(j) \tag{4}$$

2) the link utilization, denoted as U:

$$U = \sum_{j=1}^{C} j G^{-1} G(j) + \sum_{j=C+1}^{T} C G^{-1} G(j) \tag{5}$$

where $G = \sum_{j=0}^{T} G(j)$ is the normalization constant.

III. THE E-EMLM UNDER THE BR POLICY

If we apply the BR policy to the E-EMLM (E-EMLM/BR), then (3) takes the form:

$$G(j) = \begin{cases} 1 & \text{for } j = 0 \\ \frac{1}{\min(j,C)} \sum_{k \in K_e} \alpha_k D_k(j - b_k) G(j - b_k) + & \\ \frac{1}{j} \sum_{k \in K_a} \alpha_k D_k(j - b_k) G(j - b_k) & \text{for } j = 1,\ldots,T \\ 0 & \text{for } j < 0 \end{cases} \tag{6}$$

$$D_k(j - b_k) = \begin{cases} b_k & \text{for } j \le T - t(k) \\ 0 & \text{for } j > T - t(k) \end{cases} \tag{7}$$

where $t(k)$ is the reserved bandwidth (BR parameter) for service-class k calls (either elastic or adaptive).

The BR policy ensures CBP equalization among different service-classes by a proper selection of the BR parameters. If, for example, CBP equalization is required between calls of three service-classes with b_1=1, b_2=7 and b_3=10 b.u., respectively, then $t(1)$ = 9 b.u, $t(2)$=3 and $t(3)$ = 0 b.u. so that $b_1 + t(1) = b_2 + t(2) = b_3 + t(3)$. If $t(k) = 0$ for all k (k=1,…,K) then the E-EMLM results. Furthermore, if T=C then the EMLM results.

 ISBN: 978-1-61208-279-0

The application of the BR policy in the E-EMLM is based on the assumption that the number of service-class k calls is negligible in states $j > T\text{-}t(k)$ and is incorporated in (6) by the variable $D_k(j\text{-}b_k)$ given in (7). The states $j > T\text{-}t(k)$ belong to the so-called reservation space. Note that the population of calls of service-class k in the reservation space may not be negligible. In [15], [16] a complex procedure is implemented in order to take into account this population and increase the accuracy of CBP results in the EMLM and Engset multirate state-dependent loss models, respectively. However, according to [16] this procedure may not always increase the accuracy of the CBP results compared to simulation.

The CBP of service-class k, B_k, in the E-EMLM/BR is given by:

$$B_k = \sum_{j=T-b_k-t(k)+1}^{T} G^{-1}G(j) \quad (8)$$

Having obtained the values of the link occupancy distribution, $G(j)$, according to (6) we can calculate the link utilization according to (5).

Note that if the link accommodates only elastic service-classes then the link occupancy distribution can be determined by the following recursive formula [14]:

$$G(j) = \begin{cases} 1 & \text{for } j = 0 \\ \dfrac{1}{\min(j,C)} \sum_{k \in K_e} \alpha_k D_k(j-b_k) G(j-b_k) & \text{for } j = 1,...,T \\ 0 & \text{for } j < 0 \end{cases} \quad (9)$$

where the values of $D_k(j\text{-}b_k)$ are calculated by (7).

Furthermore, if the link accommodates only stream traffic (i.e., calls of all service-classes cannot compress their bandwidth) then the link occupancy distribution is given by [17]:

$$G(j) = \begin{cases} 1 & \text{for } j = 0 \\ \dfrac{1}{j} \sum_{k \in K} \alpha_k D_k(j-b_k) G(j-b_k) & \text{for } j = 1,...,C \\ 0 & \text{for } j < 0 \end{cases} \quad (10)$$

where T=C and the values of $D_k(j\text{-}b_k)$ are calculated by (7).

IV. Application Example - Evaluation

In this section, we present an application example of the new model, E-EMLM/BR, and the existing model, E-EMLM. Through the E-EMLM/BR we obtain analytical CBP and link utilization results, and we compare them with the corresponding simulation results, in order to reveal the accuracy of the proposed model. We also compare the analytical CBP and link utilization results of the E-EMLM/BR with those obtained by the E-EMLM to reveal the consistency and the necessity of the proposed model. As far as simulation results are concerned, they are mean values of 7 runs; no reliability ranges are presented, because they are very small. The simulation language used is Simscript II.5 [18].

As an application example, we consider a single link of capacity C = 100 b.u., that accommodates calls of four service-classes, with the following traffic characteristics:

1st service-class: α_1 = 12 erl, b_1 = 1 b.u.
2nd service-class: α_2 = 6 erl, b_2 = 2 b.u.
3rd service-class: α_3 = 3 erl, b_3 = 4 b.u.
4th service-class: α_4 = 2 erl, b_4 = 10 b.u.

Calls of the 1st and 2nd service-class are elastic, while calls of the 3rd and 4th service-class are adaptive. The limit T, up to which bandwidth compression of all calls is permitted, takes two different values T = 120 and T = 140 b.u. In the first case, the minimum proportion of the required peak-bandwidth takes the value: r_{min}= C/T=100/120=5/6. Similarly, in the second case: r_{min}= C/T=100/140=5/7.

When the BR policy is applied in the E-EMLM/BR, we choose the BR parameters $t(1)$=9, $t(2)$=8, $t(3)$=6 and $t(4)$=0 in order to achieve CBP equalization among the four service-classes, since: b_1+ $t(1)$=b_2+$t(2)$=b_3+$t(3)$=b_4+$t(4)$.

In the x-axis of all figures, traffic loads α_1, α_2, α_3 and α_4 increase in steps of 2, 1, 0.5 and 0.25 erl, respectively. In this way, Point 1 represents the vector (α_1, α_2, α_3, α_4) = (12.0, 6.0, 3.0, 2.0) while Point 7 is (α_1, α_2, α_3, α_4) = (24.0, 12.0, 6.0, 3.5). In Figs. 1-4, we consider the value T = 120 b.u and present the analytical and simulation CBP results of the four service-class calls, respectively, in the case of the E-EMLM/BR. For comparison, we give the corresponding analytical CBP results of the E-EMLM. The value T = 140 b.u is considered in Figs. 5-8. Due to CBP equalization achieved by the aforementioned BR parameters, the analytical CBP results of the E-EMLM/BR are exactly the same in Figs. 1-4. The same happens in Figs. 5-8. The differences between the CBP results of the E-EMLM/BR and the E-EMLM show the consistency and the necessity of the E-EMLM/BR. Finally, in Fig. 9, we consider both values of T and present the analytical and simulation results of the link utilization in the case of the E-EMLM/BR. For comparison, we give the corresponding analytical results for the E-EMLM.

All figures show that the analytical and simulation CBP results of the E-EMLM/BR are very close; this fact reveals the accuracy of the E-EMLM/BR. In addition, the comparison of Fig. 1 with Fig. 5, Fig. 2 with Fig. 6, Fig. 3 with Fig. 7 and Fig. 4 with Fig. 8 shows that the increase of T reduces the values of CBP of each service-class; this fact shows the consistency of the E-EMLM/BR. Furthermore, note that the choice of the BR parameters that achieve CBP equalization, actually favors calls of the 4th service-class only (see Figs 1-4 and Figs 5-8, where the CBP of the first three service-classes increase in the E-EMLM/BR compared to the corresponding CBP results in the E-EMLM). As far as the link utilization results of Fig. 9 are concerned, they show that the compression/expansion mechanism increases the link utilization, since it decreases CBP.

 ISBN: 978-1-61208-279-0

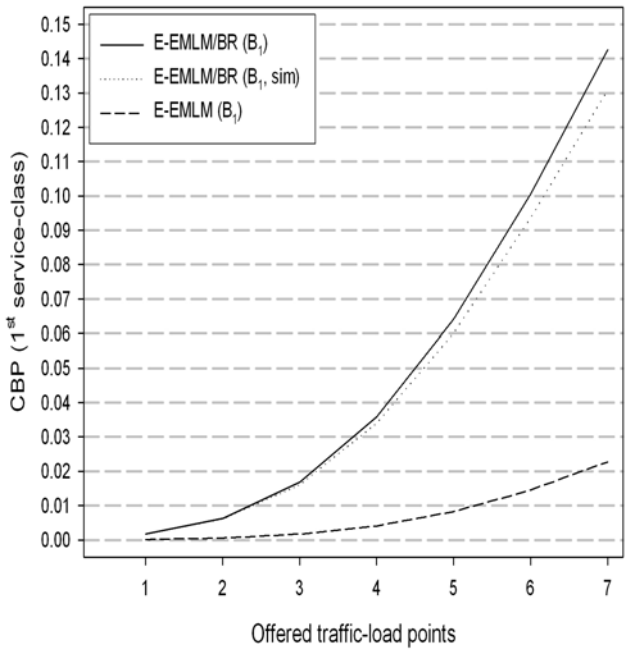


Figure 1. CBP of the 1st service-class (E-EMLM, E-EMLM/BR, *T*=120).

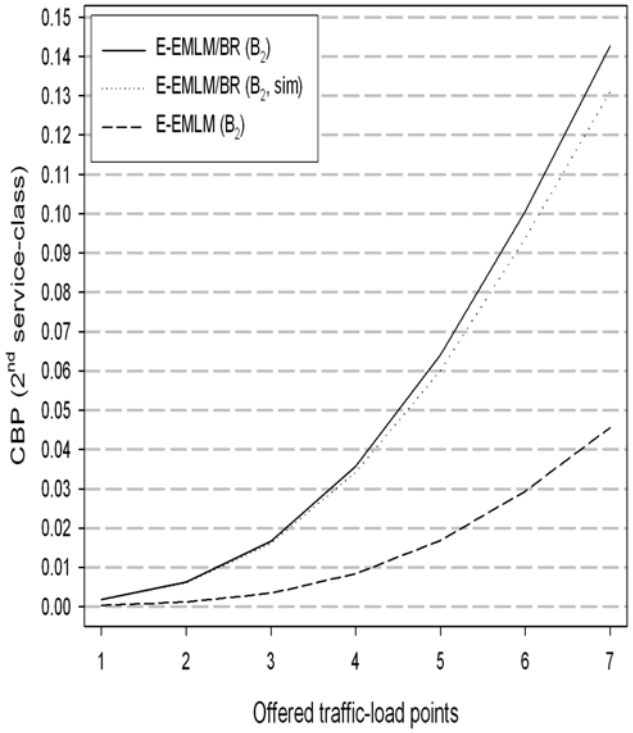


Figure 2. CBP of the 2nd service-class (E-EMLM, E-EMLM/BR, *T*=120).

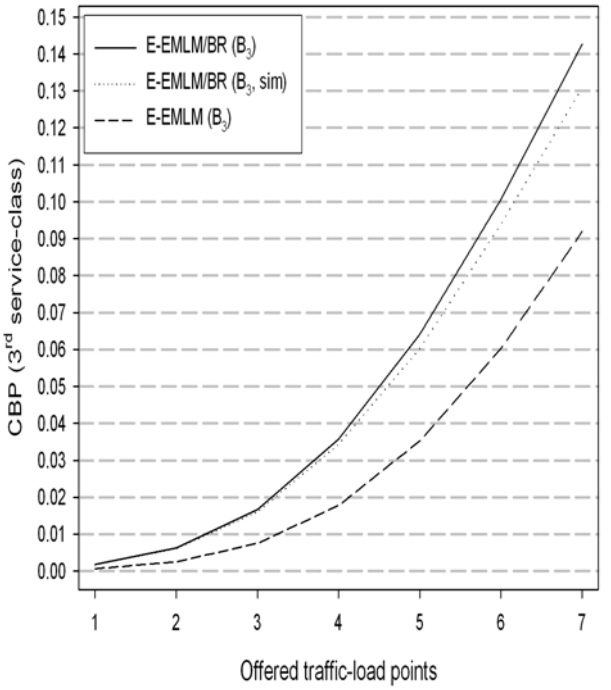


Figure 3. CBP of the 3rd service-class (E-EMLM, E-EMLM/BR, *T*=120).

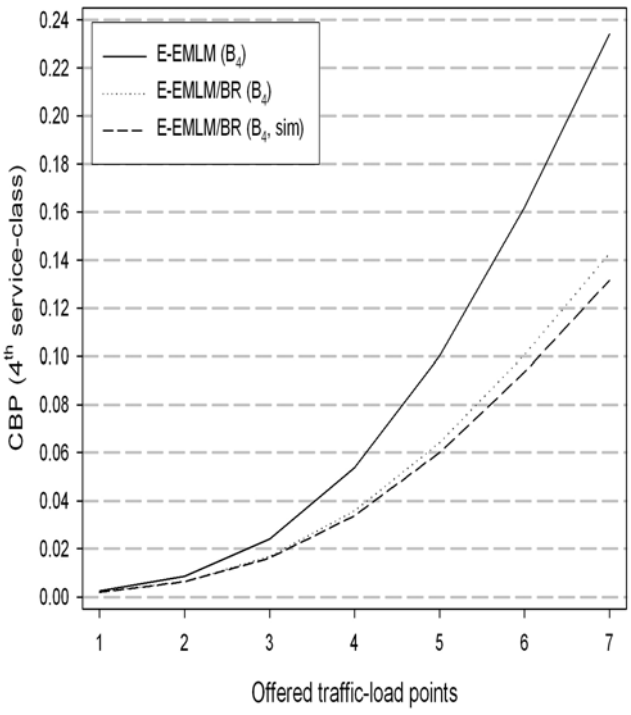


Figure 4. CBP of the 4th service-class (E-EMLM, E-EMLM/BR, *T*=120).

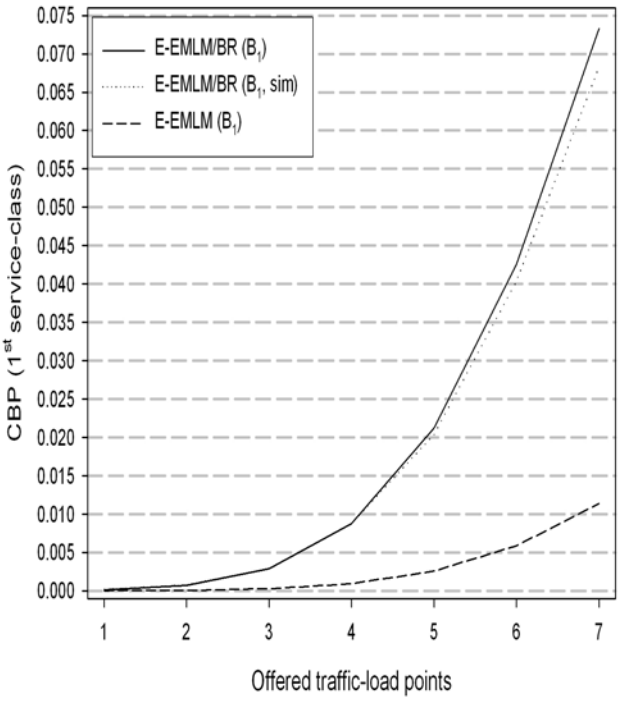


Figure 5. CBP of the 1st service-class (E-EMLM, E-EMLM/BR, *T*=140).

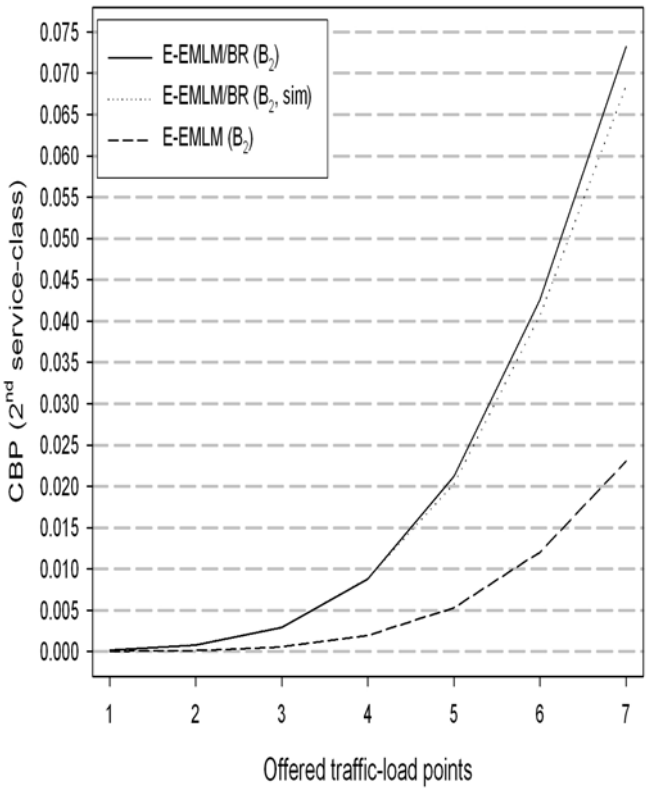


Figure 6. CBP of the 2nd service-class (E-EMLM, E-EMLM/BR, *T*=140).

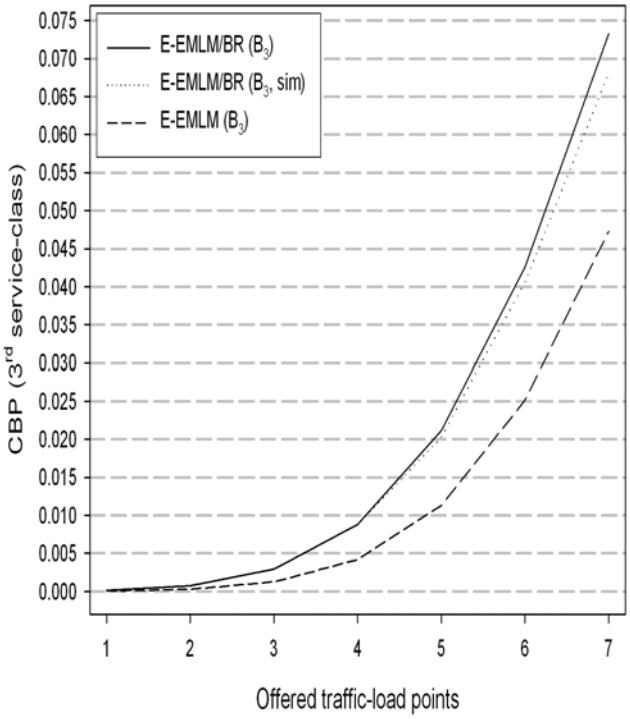


Figure 7. CBP of the 3rd service-class (E-EMLM, E-EMLM/BR, *T*=140).

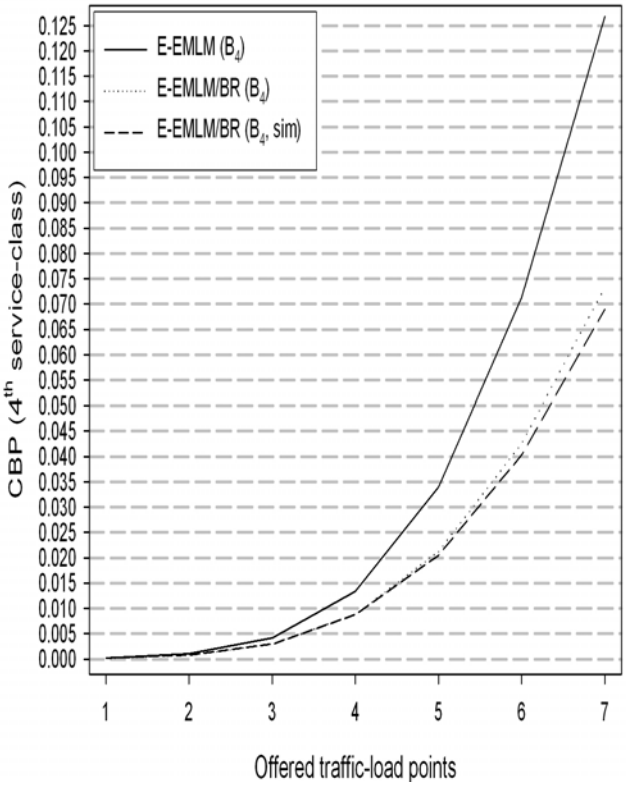


Figure 8. CBP of the 4th service-class (E-EMLM, E-EMLM/BR, *T*=140).

 ISBN: 978-1-61208-279-0

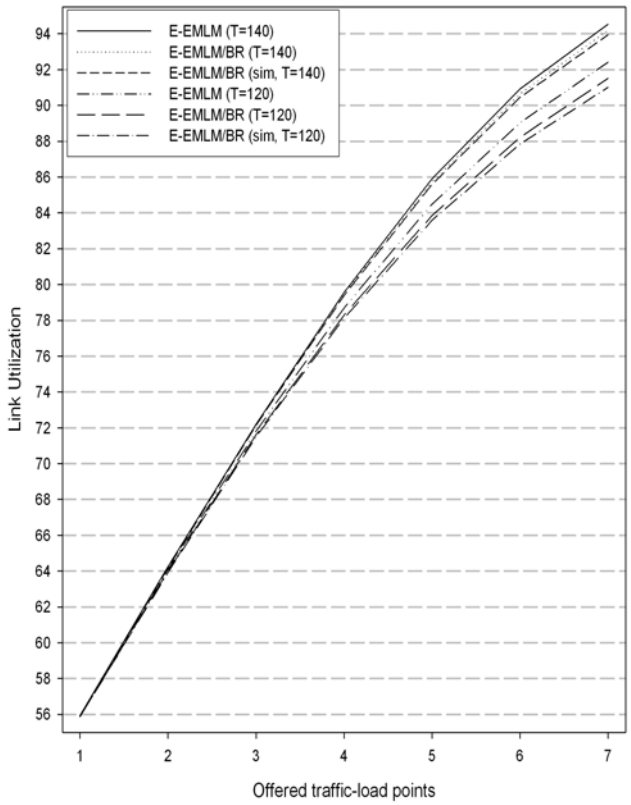


Figure 9. Link Utilization (E-EMLM, E-EMLM/BR, *T*=120 and *T*=140).

V. CONCLUSION

We propose an analytical model for the recursive calculation of CBP and link utilization in a single communication link, which accommodates multirate traffic of elastic and adaptive calls that follow a Poisson process, under the bandwidth reservation policy. This policy is used in order to achieve certain QoS among elastic and adaptive calls, or obtain equalization of call blocking probabilities. Simulation results verify the analytical CBP and link utilization results, and prove the accuracy of the proposed model. The comparison between the analytical results of the proposed model (E-EMLM/BR) and the corresponding results of the existing model (E-EMLM), prove the necessity and the consistency of the new model. A future extension of this model is the consideration of a mixture of random (Poisson) and quasi-random traffic under the BR policy. Quasi-random traffic, i.e., traffic generated by a finite number of sources (in contrast to random traffic, which assumes an infinite number of traffic sources), is a necessary assumption to many realistic network configurations.

REFERENCES

[1] G. Stamatelos and V. Koukoulidis, "Reservation – Based Bandwidth Allocation in a Radio ATM Network", IEEE/ACM Trans. Networking, vol. 5, June 1997, pp.420-428.

[2] S. Rácz, B. Gerő, and G. Fodor, "Flow level performance analysis of a multi-service system supporting elastic and adaptive services", Performance Evaluation, vol. 49, Sept. 2002, pp. 451-469.

[3] V. Vassilakis, I. Moscholios, M. Logothetis, "Evaluation of Multi-rate Loss Models for Elastic Traffic", 3rd International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs'05), Ilkley, West Yorkshire, U.K., July 18-20, 2005.

[4] V. Vassilakis, I. Moscholios, M. Logothetis, "Evaluation of Multi-rate Loss Models for Elastic and Adaptive Services", 12th Polish Teletraffic Symposium (PSRT '05), Poznan, Poland, September 19-20, 2005.

[5] V. Vassilakis, I. Moscholios and M. Logothetis, "The Extended Connection Dependent Threshold Model for Elastic and Adaptive Traffic", Proc. of Communication Systems, Networks and Digital Signal Processing – 5th CSNDSP' 2006, Patras, Greece, 19-21 July 2006.

[6] G. Fodor and M. Telek, "Bounding the Blocking Probabilities in Multirate CDMA Networks Supporting Elastic Services", IEEE/ACM Trans. on Networking, vol. 15, Aug. 2007, pp.944-956.

[7] V. Vassilakis, G. Kallos, I. Moscholios, and M. Logothetis, "Call-Level Analysis of W-CDMA Networks Supporting Elastic Services of Finite Population", IEEE ICC, May 2008, pp.285-290.

[8] I. Moscholios, J. Vardakas, M. Logothetis and A. Boucouvalas, "A Batched Poisson Multirate Loss Model Supporting Elastic Traffic under the Bandwidth Reservation Policy", Proc. of IEEE International Conference on Communications, ICC 2011, Kyoto, Japan, 5-9 June 2011.

[9] I. Moscholios, J. Vardakas, M. Logothetis and A. Boucouvalas, "QoS Guarantee in a Batched Poisson Multirate Loss Model Supporting Elastic and Adaptive Traffic", Proc. of IEEE ICC 2012, Ottawa, Canada, 10-15 June 2012.

[10] I. Moscholios, V. Vassilakis, J. Vardakas and M. Logothetis, "Call Blocking Probabilities of Elastic and Adaptive Traffic with Retrials", Proc. of 8th Advanced Int. Conf. on Telecommunications, AICT 2012, Stuttgart, Germany, 27 May-1 June 2012.

[11] B. Gerő, P. Pályi and S. Rácz, "Flow-level performance analysis of a multi-rate system supporting stream and elastic services", Int. J. Commun. Syst., Wiley, doi: 10.1002/dac.1383, 2012.

[12] J. Kaufman, "Blocking in a shared resource environment", IEEE Trans. Commun. vol. 29, Oct. 1981, pp. 1474-1481.

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

[14] I. Moscholios, V. Vassilakis, M. Logothetis and A. Boucouvalas, "Blocking Equalization in the Erlang Multirate Loss Model for Elastic Traffic", Proc. of 2nd Int. Conference on Emerging Network Intelligence, EMERGING 2010, Florence, Italy, 25-30 October 2010.

[15] M. Stasiak and M. Glabowski, "A simple approximation of the link model with reservation by a one-dimensional Markov chain", Performance Evaluation, vol. 41, July 2000, pp.195–208.

[16] I. Moscholios and M. Logothetis, "Engset multi-rate state-dependent loss models with QoS guarantee", Int. J. Commun. Syst., vol. 19, February 2006, pp.67-93.

[17] J. Roberts, "Teletraffic models for the Telecom 1 Integrated Services Network", Proceedings of ITC-10, Mondreal, Canada, 1983.

[18] Simscript II.5, http://www.simscript.com

 ISBN: 978-1-61208-279-0

Efficiency Evaluation of Shortest Path Algorithms

Mariusz Głąbowski, Bartosz Musznicki, Przemysław Nowak and Piotr Zwierzykowski
Poznan University of Technology, Faculty of Electronics and Telecommunications
Chair of Communication and Computer Networks, Poznan, Poland
bartosz@musznicki.com, przemyslaw.nowak@inbox.com

***Abstract*—While the ever growing computational capabilities of devices that are used for man-machine interaction are taken for granted, the need to find their most optimum use is as important as ever. This issue is particularly relevant when considering solutions where the determination of the shortest path between given points (nodes) is one of the basic operations. In more complex executions of the shortest paths, sets of paths with the shortest distance between a single initial (source) point and all other destination points, as well as between all pairs of points, are to be found. For each of these approaches, individual algorithms with specific features have been worked out over the past decades. With that in mind, the present article seeks to explore this problem and is structured in such a way as to describe some of the selected algorithms solving the shortest path problem, and to analyse the efficiency of these algorithms during their operation in directed graphs of different type. The study shows that the efficiency varies among algorithms under investigation and allows to suggest which one ought to be used to solve a specific variant of the shortest path problem.**

Keywords-shortest path; algorithms; efficiency; evaluation

I. INTRODUCTION

The foundations for the present evaluation of the algorithms presented in this article are given by the research studies on shortest path problem solving using Ant Colony Optimization (ACO) metaheuristic approach [1]. It is just in the initial stage in the assessment of the potential in the applications of the ACO algorithm that the authors decided to start an in-depth analysis of those algorithms that represented a more traditional approach to the problem. As a result of the following studies, relevant tests have been carried out which are to be presented and compared in this article. It should be stressed that both well-known [2] and less commonly used algorithms are presented as long as they provide a possibility of finding the optimal solution having first satisfied some pre-defined initial requirements. Heuristic ACO algorithms have not been included in the presented evaluation for the simple reason that their operation does not, in fact, guarantee finding a solution that would always be optimal [3]. Moreover, the results obtained on the basis of ACO can be strongly dependent on the structure of the graph and there is no guarantee that any solution of any kind would be found at all [1].

In the process of careful investigation of publications related to the shortest path problem numerous books and papers have been studied. The most of comparison papers are either directed at specific aspects and applications of the algorithms [4]–[6] or are focused on comparing new concepts with more classical methods [7], [8]. Some papers are concerned with asymptotic computational complexity [9]–[12] while other works are aimed at empirical computational complexity analysis of a number of algorithms based on implementation and simulation [5], [13]–[16]. In this paper, we decided to follow the latter approach to build this article upon experimental findings with respect to practical performance of a range of 12 closed-form complexity algorithms for solving shortest path problems. The introduced homogeneous data structure representing graphs under scrutiny is carefully discussed. Owing to the well-defined data structure, the results can be directly compared what is critical to conclusively evaluate the efficiency.

The contents of the subsequent sections are arranged as follows. Section II shows the problem of the shortest path and lists some of its applications. The relevance to and relationship with the shortest path tree is discussed in Section III. In addition, a description of the two groups of algorithms that have been put to the analysis is presented. The data structure that represents the graphs under consideration is discussed in Section IV. Later on, in Section V, the graphs in which simulations were carried out are described. The description is followed by Section VI that will focus on the presentation and discussion of the results of the study. Finally, in Section VII, the article is summed up with conclusions.

II. PROBLEM OF THE SHORTEST PATH

For the directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, where N is the set of nodes (vertices) and A is the set of arcs (edges), we assign the cost a_{ij} to each of its edges $(i, j) \in \mathcal{A}$ (alternatively, this cost can be also called the length). We denote the biggest absolute value of an edge cost by C. For the resulting path $(n_1, n_2, \ldots, n_k)$, its length can be expressed by (1).

$$a_{ij} = \sum_{i=1}^{k-1} a_{n_i n_{i+1}} \tag{1}$$

A path is called the shortest path if it has the shortest length from among all paths that begin and terminate in given vertices. The shortest path problem involves finding paths with shortest lengths between selected pairs of vertices. The

 ISBN: 978-1-61208-279-0

initial vertex will be designated as s, while the end vertex as t.

A number of basic variants of the shortest path problem can be distinguished [17]:

- finding the shortest path between a pair of vertices.
- finding the shortest paths with single initial vertex.
- finding the shortest paths with single end vertex.
- finding the shortest paths between all pairs of vertices.

In solving the problem of the shortest path we shall apply the following assumptions (which, in the case of some specific algorithms, may not be required).

- *Costs of the edge a_{ij} are integers* (this requirement applies to only some of the algorithms). In the case of the real costs of the edge, we can convert summations to integers multiplying them by an appropriately high number. Imaginary values would introduce unnecessary complications with their representations in computer-mediated activities.
- *There is a directed path between the pairs of vertices under consideration.*
- *The graph does not include negative cycles.* The problem of the shortest path with negative cycles is $\mathcal{NP}$-hard (impossible to be presented using a polynomial algorithm).
- *The graph is a directed graph.* In the case of the undirected graph with non-negative weights, it is easy to to transform it into a directed graph.

The solution for the problem of the shortest path finds its application in a number of areas such as transportation or routing in communication networks [2], [18], [19] and is often related to searching for the shortest path tree in a graph.

III. Algorithms for solving Shortest Path problems

The following subsections of this Section focus on the algorithms for a determination of the shortest paths between a given single initial vertex and all the remaining vertices of the graph. It can be proved that the shortest paths from one node of a graph to all of the remaining nodes create a shortest paths tree [17], [20]. A characteristic feature of this tree is the fact that its root is formed from the initial (source) vertex, all of its edges are directed in the direction opposite to the vertex, and each path that can be created from the initial vertex to any other vertex is the shortest path to this vertex.

The algorithms solving shortest path problems that are briefly discussed in the following subsections have been evaluated through an efficiency analysis. Each of the algorithms has particular features that eventually lead to their differences in their properties and performance. On account of their possible applications, the algorithms have been, in turn, divided into two categories.

A. Single-Source Shortest Paths problem

The following subsections of this Section focus on the algorithms for a determination of the shortest paths between a given single initial vertex and all the remaining vertices of the graph.

1) Generic algorithm: The operation of the generic algorithm [21] is based on iterative checking of edges from the vertex under consideration i and on label setting for vertex j, in which a given edge terminates, to $d_j = d_i + a_{ij}$, in the case when $d_j > d_i + a_{ij}$. To store the vertices that are to be checked, the list V is used, called *candidates list*. The way vertices are stored in this list, as well as the method determining the addition and the retrieval of vertices to and from it, is frequently the major factor that distinguishes individual algorithms under consideration. In the case of the generic algorithm, the candidates list is a *FIFO* queue in which operations of additions and retrieval of a vertex to the end of it or from its head, respectively, are performed.

2) Dijkstra's algorithm: Dijkstra's algorithm is presumably the best known algorithm for finding the shortest path in the directed graph [22]. The basic difference between this algorithm and the generic algorithm is the way in which vertices are drawn from the candidates list — the selected vertex is the vertex that has the smallest label from all available vertices in the list:

$$d_i = \min_{j \in V} d_j \tag{2}$$

This causes the vertex with its label set, as well as all vertices that are in the path from the initial vertex to this particular vertex, to have the minimum value of the label and to not be added again to the candidates list. The total number of operations that the Dijkstra's algorithm needs to perform to solve the shortest path problem is $O(N^2)$.

3) Dijkstra's algorithm using a heap: It is not possible to decrease the number of operations that are performed in order to check labels, because this would not make it possible to guarantee the optimal solution finding — each edge has to be checked at least once. A selection of an optimal data structure that represents the list of candidates makes it possible, in turn, to reduce significantly the computational complexity of the operation of the selection of a vertex from the candidates list [23]. Here, heaps (also known as priority queues) can serve ideally the purpose. Using Fibonacci heap we can solve the shortest path problem using Dijkstra's algorithm and performing $O(A + N \log N)$ operations.

4) Dial's algorithm: Another way to reduce the number of operations accompanying the selection of a vertex from the candidates list is a division of the list into buckets [24]. Each bucket B_k stores only vertices with a given label k. This causes lengths of edges to have to be integers and non-negative. The computational complexity of the Dial's algorithm is $O(A + NC)$. What is crucial to understand, is that the bucket deletion and insertion operations require

 ISBN: 978-1-61208-279-0

linear time and not more than NC buckets need to be examined by the procedure [14]. The higher the absolute value of an arc cost C, the more operations need to be performed by the algorithm, and thus, the performance gain related to the usage of buckets dramatically diminishes. Therefore, for small values $C \ll N$, Dial's algorithm performs very well in practice.

5) Bellman-Ford algorithm: The Bellman-Ford algorithm belongs to algorithms of the *label-correcting* type that treat all labels for vertex distances as temporary until the last iteration, after which all labels are set to optimal values [25]. This algorithm provides a possibility to solve the shortest path problem in graphs with negative lengths of edges. In the case when a negative cycle is found, the algorithm yields falsehood as the result of its operation. This algorithm makes $N - 1$ iterations in which it checks A edges. Its computational complexity is then equal to $O(NA)$.

6) D'Esopo-Pape algorithm: The D'Esopo-Pape algorithm uses the candidates list in the form of a queue [26]. Vertices that are to be checked are always retrieved from the head of the list. However, the place a given vertex is added to in the candidate list depends on whether the vertex has already been located in this list. If so, it is added to its head, otherwise — to the end of the list.

7) SLF algorithm: The Small Label First algorithm (*SLF*) seeks to manage the candidates list in such a way as to make vertices with small labels located as close to the head of the list as possible [27]. The reason for this operation is the fact that the smaller the label of a vertex that is retrieved from the candidates list, the lower the probability that this vertex will be forwarded to the list once again. This algorithm, just as the two following algorithms, attempts to reach the characteristic operation of Dijkstra's algorithm with a lower computational outlay.

8) LLL algorithm: The Large Label Last algorithm (*LLL*) attempts to achieve the operation that is similar to that of the previous algorithm using a specific method for the retrieval of vertices from the candidates list [28]. The addition of vertices to the candidates list is not defined in any way. However, the method for their retrieval from the list is defined. Each time when a vertex is to be taken from the list, the average value of the labels of the vertices in the list is calculated. Then, the label of the vertex that is at the head of the list is compared with this average. If the label of the vertex is higher than the average, the vertex is moved to the end of the list. Otherwise, the vertex is returned as the one that has to be considered in this iteration.

9) SLF/LLL algorithm: The *SLF*/*LLL* combines the *SLF* algorithm method for the addition of vertices to the candidates list and the *LL* algorithm method for their retrieval from the list [21]. The *SLF*/*LLL* algorithm requires a lower number of iterations to solve the shortest path problem than the algorithms it combines. This is done, however, at the cost of the increased number of necessary calculations.

B. All-Pairs Shortest Path problem

The following subsections present algorithms that are dedicated to finding the shortest paths between all pairs of vertices.

1) The doubling algorithm: Algorithm's operation is based on iterative calculation of the shortest paths for all vertices composed of an increasing number of edges [29]. It starts with paths that are composed of just one edge, and then checks whether paths that are composed of two edges would not be shorter. This operation is then repeated until all paths that are composed of $N-1$ edges are checked. Bearing in mind the fact that a path that is composed of more than $N - 1$ edges cannot be shorter than the shortest path, we know that $D^n = D^{(N-1)}$ for all $n \geq N - 1$. This gives the ultimate computational complexity of the algorithm equal to $\Theta(n^3 \log_2 N)$.

2) Floyd-Warshall algorithm: The Floyd-Warshall algorithm obtains what the previous algorithm was capable of, using a different approach and achieving at the same time lower computational complexity equal to $\Theta(N^3)$ [30], [31].

3) Johnson algorithm: For sparse graphs (i.e., those in which the number of edges is far lower than N^2) it is possible to improve the process of calculation of the shortest paths between all pairs of vertices using Johnson algorithm [32]. For this purpose, the two algorithms discussed earlier, i.e., the Bellman-Ford algorithm and Dijkstra's algorithm (most favourably in its form with a heap), are used. If we choose to apply the implementation of Dijkstra's algorithm with Fibonacci heap, then we are obliged to perform $O(NA + N^2 \log N)$ operations to calculate the shortest paths between all the pairs of vertices in a sparse graph. Using a binary heap would result in an increase in the number of necessary operations to $O(NA \log N)$.

IV. Data structure representing graphs

To represent graphs during the simulation, a double associative adjacency array was used. This structure is composed of two associative arrays — one (external), representing vertices from which edges originate, and the other (internal) representing all vertices which edges for a given row of the first matrix (table) join. Such a representation provides an opportunity to minimize shortcomings of typical structures,

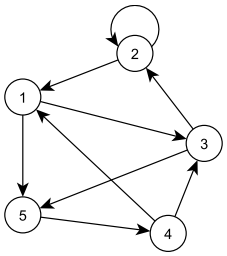


Figure 1. Exemplary directed graph

 ISBN: 978-1-61208-279-0

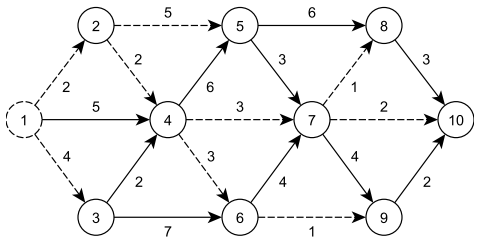


Figure 2. Manually created *custom* graph in which the edges marked with the dashed line create a shortest paths tree with the root in node 1

such as the list of edges or the adjacency matrix, providing at the same time appropriately low computational complexity for individual operations. The applied structure makes it possible to store additional information about edges, e.g., weights or costs. A homogeneous method for the projection (mapping) of graphs for all simulated algorithms ensures further comparability of the results of simulations.

The operation of the structure may differ depending on the implementation of the associative array and is dependable on the programming language used if embedded structures are used. The most crucial operation is the operation of checking whether a given key is in the array, hence structures that handle this best, e.g., hash tables or self-balancing binary search trees, are applied. Additionally, we can adjust the operation of the double associative adjacency array for our particular needs and thus make it possible, for example, to sort vertices in the internal array which a given edge joins using a heap.

For a graph with edge weights, the double, associative adjacency array T_{2asoc} can be written as follows:

$T_{2asoc} = T_{ext}$	external array
$T_{2asoc}[i] = T_{ext}[i] = T_{int_i}$	internal array for edge coming out from vertex i
$T_{2asoc}[i][j] = T_{int_i}[j] = a_{i,j}$	edge weight (i, j)

For example, the graph in Fig. 1 will be mapped in the following way:

$$
\begin{array}{lll}
T_{2asoc}[1] & = T_{int_1} & \\
T_{2asoc}[1][3] & = T_{int_1}[3] & = a_{1,3} \\
T_{2asoc}[1][5] & = T_{int_1}[5] & = a_{1,5} \\
T_{2asoc}[2] & = T_{int_2} & \\
T_{2asoc}[2][1] & = T_{int_2}[1] & = a_{2,1} \\
T_{2asoc}[2][2] & = T_{int_2}[2] & = a_{2,2} \\
T_{2asoc}[3] & = T_{int_3} & \\
T_{2asoc}[3][2] & = T_{int_3}[2] & = a_{3,2} \\
T_{2asoc}[3][5] & = T_{int_3}[5] & = a_{3,5} \\
T_{2asoc}[4] & = T_{int_4} & \\
T_{2asoc}[4][1] & = T_{int_4}[1] & = a_{4,1} \\
T_{2asoc}[4][3] & = T_{int_4}[3] & = a_{4,3} \\
T_{2asoc}[5] & = T_{int_5} & \\
T_{2asoc}[5][4] & = T_{int_5}[4] & = a_{5,4}
\end{array}
$$

Characteristic features of the structure:

- required memory: $O(N + A)$
- effective memory complexity for directed sparse graphs
- effective execution of graph algorithms that require to reach all vertices adjacent to a given vertex (logarithmic complexity)
- capacity of remembering parallel edges
- effective execution of checking whether the graph includes a given edge (logarithmic complexity)
- effective execution of addition and removal of edges of a graph (logarithmic complexity)
- possibility of a substitution of the internal associative table with some other structure, e.g., in order to sort vertices in which a given edge terminates by the weight of the edge (e.g., using binary, Fibonnaci, binomial or Relaxed heap)
- fairly complicated in its execution

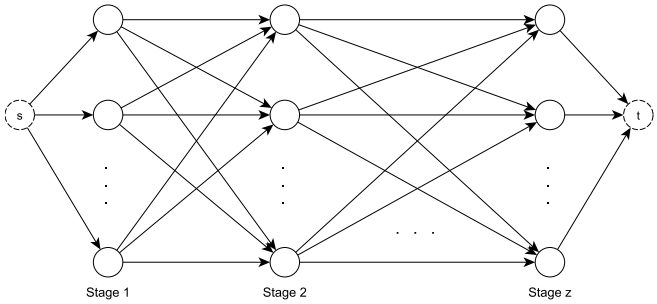


Figure 3. A graph that presents the problem of the shortest path in a multi-stage graph

TABLE I. STRUCTURE OF THE GRAPHS USED IN THE SIMULATION

graph	vertices	edges	
		number	lengths
custom	10	19	$\langle 1, 7 \rangle$
multistage	52	420	$\langle 1, 9 \rangle$
random	25	125	$\langle 1, 9 \rangle$

V. GRAPHS USED IN THE SIMULATION

To examine the efficiency and performance of the algorithms during their operation in different graphs, directed graphs constructed manually and those that were generated pseudo-randomly were used. To discuss the results, the 3 representative graphs described in Table I were selected. Graph *custom* shown in Fig. 2 was created manually from 10 vertices that were joined together by 19 edges.

Another graph that was used in the tests is the graph that is characteristic for a multi-stage shortest path problem. An exemplary graph is presented in Fig. 3. The multi-stage graph used in the tests, *multistage*, has 5 stages, each having 10 vertices. The lengths of edges were generated randomly from within the interval $\langle 1, 9 \rangle$. The *random* graph was generated randomly, without loops, and with 5 edges coming out of each of the vertices.

 ISBN: 978-1-61208-279-0

VI. RESULTS OF THE SIMULATIONS OF THE ALGORITHMS

All the tests were carried out in a simulation environment prepared in C# programming language. In order to achieve reliable results, each algorithm was performed 100 times for each of the graphs. To eliminate the influence of the simulation environment, extreme results were rejected and then the average of the remaining results was calculated.

Table II shows the running times of the algorithms tested for the graphs discussed in Section V. The results are divided into two groups — algorithms solving Single-Source Shortest Paths problem (*SSSP*) and algorithms solving All-Pairs Shortest Path problem (*APSP*).

The graph *custom* was solved by all *SSSP* algorithms in almost identical times. Of all the algorithms only two deserve a mention here — Dijkstra's algorithm with a heap (that operated within the longest time), and *SLF* (that solved the problem slightly quicker than the rest). The results that were very similar to that of the *SLF* algorithm were also shared by Dijkstra's algorithm, Dial's algorithm and the *LLL* algorithm. From the group of the *APSP* algorithms, it was the Floyd-Warshall algorithm that fared the best, being less than twice as long as the *SSSP* algorithms. The remaining algorithms needed about twice as much time to find all paths.

The graph characteristic for the multi-stage shortest path problem (*multistage*) brought a significant increase in differences between *SSSP* algorithms. Again, the *SLF* algorithm was the quickest, whereas Bellman-Ford and D'Esopo-Pape algorithms handled the problem the worst. Except Dijkstra's algorithm with a heap, which was performing slightly longer than the rest, the remaining algorithms had similar running times. This situation for the *APSP* algorithms was different than in the case of the previous graph — Johnson algorithm was the quickest and the doubling algorithm was the slowest.

The last graph under consideration (*random*) was solved the quickest in the *SSSP* mode by the *SLF* algorithm, with

TABLE II. COMPARISON OF RUNNING TIMES FOR THE ALGORITHMS SOLVING THE SHORTEST PATH PROBLEM IN MICROSECONDS

group	algorithm	graph		
		custom	multistage	random
SSSP	generic	112	312	163
	Dijkstra	100	324	148
	DijkstraHeap	146	466	200
	Dial	104	322	172
	Bellman-Ford	119	3252	511
	D'Esopo-Pape	113	1260	239
	SLF	96	262	143
	LLL	102	336	155
	SLF/LLL	112	318	161
APSP	doubling algorithm	324	47678	4756
	Floyd-Warshall	184	16045	2057
	Johnson	418	9309	2959

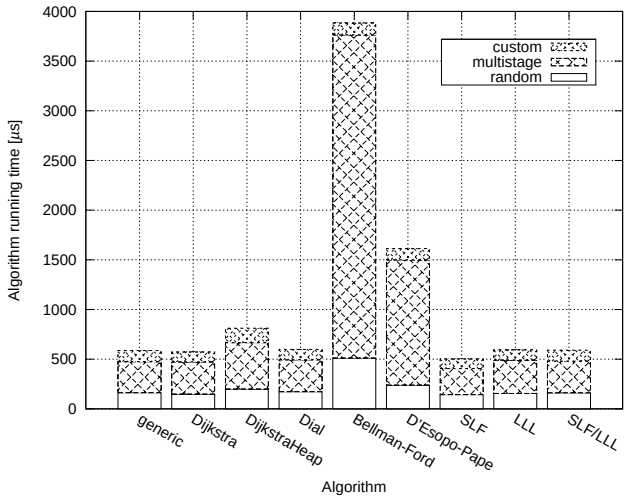


Figure 4. Chart of aggravated running times of the algorithms solving the shortest path problem with one initial vertex (*SSSP*)

Dijkstra's algorithm as the runner up and the Bellman-Ford and the D'Esopo-Pape algorithms well behind the two. The latter two were the worst as compared to all involved *SSSP* algorithms. This time, the quickest *APSP* algorithm was the Floyd-Warshall algorithm. Johnson algorithm performed slightly worse, while the doubling algorithm was the worst (the longest) of the lot.

The procedures that solve the *SSSP* problem best include the *SLF* algorithm, that had the shortest times for each tested graph, and Dijkstra's algorithm, that always performed with a quite similar time. The *LLL* and the *SLF/LLL* algorithms performed very well and did not generate solutions over times that differ much from those provided by the quickest algorithm. The generic algorithm and Dial's algorithm performed slightly better or slightly worse depending on the chosen graph. Dijkstra's algorithm with a heap had some problems and, instead of performing quicker than Dijkstra's algorithm, was slower. In this particular case, this can be most probably explained by the missing optimization of the heap that formed the base for the algorithm. Undoubtedly, however, an improvement in the running time during which solutions are provided is still possible. At least, an improvement in the running time needed for the algorithm to generate solutions is possible. As it is clear from Fig. 4, for the Bellman-Ford and D'Esopo-Pape algorithms, the worst case occurs far too often, which may result from both non-optimal implementation and from the possibility of their operation on graphs that were unsuitable for them. The D'Esopo-Pape algorithm was much quicker to solve graphs, but irrespective of the fact it underperformed far too much as compared to the rest of the algorithms. Underperformance of the latter group of algorithms is particularly visible in graphs that have a higher number of edges, which results from the assumptions, as they were, that served as a basis

 ISBN: 978-1-61208-279-0

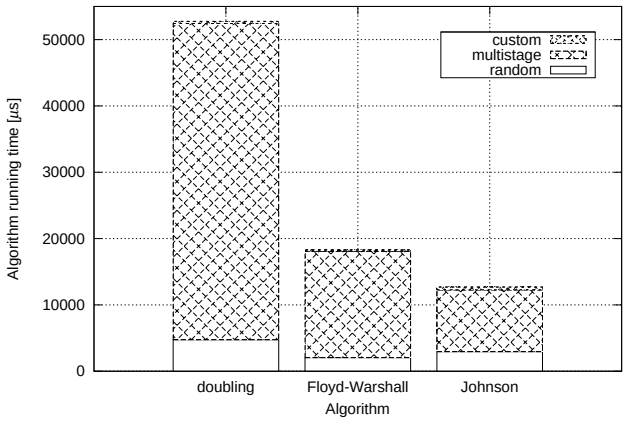


Figure 5. Chart of aggravated running times of the algorithms solving the shortest path problem between all pairs of vertices (*APSP*)

for their design.

The *APSP* algorithms were decidedly varied across different performance dimensions, in particular in relation to the time necessary to generate results, which is clearly shown in Fig. 5. The doubling algorithm was the slowest and performed several times slower than the competitors. The Floyd-Warshall algorithm was the fastest for 2 graphs, while for the third graph it was in second place. The differences in the time needed for graphs to be solved are in its case significant as compared to Johnson algorithm that overall turned out to be the fastest one.

VII. Conclusion

This article presented 12 algorithms solving the shortest path problem and provided an evaluation of their efficiency. The study showed that in a prepared simulation environment that ensured directed graphs of different type to be provided, the weakest aggregated time results from among all the available algorithms solving the Single-Source Shortest Paths problem were those of, in the descending order, the Bellman-Ford and the D'Esopo-Pape algorithms. The fastest algorithm was Small Label First algorithm, slightly faring better than Large Label Last algorithm. From the pool of the algorithms dedicated for All-Pairs Shortest Path problem, the doubling algorithm performed decidedly worst, while the best results were those of Floyd-Warshall algorithm (two graphs) and Johnson algorithm (one graph).

In addition to the presentation of run-time relationships between the algorithms, the study indicated the importance and significance of an appropriate choice of a method destined to solve the problem that would be the most efficient for a type of the graph structure to be used. Moreover, details concerning the implementation as well as the architecture of the structures for the representation of data can significantly influence the performance of an algorithm.

References

[1] M. Głąbowski, B. Musznicki, P. Nowak, and P. Zwierzykowski, "Shortest Path Problem Solving Based on Ant Colony Optimization Metaheuristic," International Journal of Image Processing & Communications, Special Issue: Algorithms and Protocols in Packet Networks, vol. 17, no. 1–2, 2012, pp. 7–17.

[2] B. Y. Wu and K.-M. Chao, Spanning Trees and Optimization Problems. USA: Chapman & Hall/CRC Press, 2004.

[3] C. Blum and A. Roli, "Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison," ACM Computing Surveys, vol. 35, no. 3, September 2003, pp. 268–308.

[4] R. Vasappanavara, E. V. Prasad, and M. N. Seetharamanath, "Comparative Studies of Shortest Path Algorithms and Computation of Optimum Diameter in Multi Connected Distributed Loop Networks," Multi-, Inter-, and Transdisciplinary Issues in Computer Science and Engineering, vol. 2, no. 1, January 2006, pp. 62–67.

[5] B. V. Cherkassky, L. Georgiadis, A. V. Goldberg, R. E. Tarjan, and R. F. Werneck, "Shortest Path Feasibility Algorithms: An Experimental Evaluation," ACM Journal of Experimental Algorithmics, vol. 14, 2009.

[6] K. Gutenschwager, A. Radtke, S. Völker, and G. Zeller, "The Shortest Path - Comparison of Different Approaches and Implementations for the Automatic Routing of Vehicles," in Proceedings of the 2012 Winter Simulation Conference, Berlin, Germany, 9–12 December 2012.

[7] U. Lauther, "An Experimental Evaluation of Point-To-Point Shortest Path Calculation on Road Networks with Precalculated Edge-Flags," in Proceedings of Ninth DIMACS Implementation Challenge, Piscataway, NJ, USA, 13–14 November 2006.

[8] Y. Sharma, S. C. Saini, and M. Bhandhari, "Comparison of Dijkstra's Shortest Path Algorithm with Genetic Algorithm for Static and Dynamic Routing Network," International Journal of Electronics and Computer Science Engineering, vol. 1, no. 2, 2012, pp. 416–425.

[9] S. Pettie, "On the Comparison-Addition Complexity of All-Pairs Shortest Paths," in Proceedings of ISAAC 2002, 13th International Symposium on Algorithms and Computation, Vancouver, BC, Canada, 21–23 November 2002.

[10] J. Hershberger, S. Suri, and A. Bhosle, "On the Difficulty of Some Shortest Path Problems," ACM Transactions on Algorithms, vol. 3, no. 1, 2007.

[11] R. Cohen and G. Nakibly, "On the Computational Complexity and Effectiveness of N-hub Shortest Path Routing," IEEE/ACM Transactions on Networking, vol. 16, no. 3, 2008, pp. 691–704.

[12] L. Roditty and U. Zwick, "On Dynamic Shortest Paths Problems," Algorithmica, vol. 61, no. 2, 2011, pp. 389–401.

 ISBN: 978-1-61208-279-0

[13] B. L. Golden, “Shortest Path Algorithms: A Comparison,” Massachusetts Institute of Technology, Operations Research Center, Tech. Rep., October 1975.

[14] B. V. Cherkassky, A. V. Goldberg, and T. Radzik, “Shortest paths algorithms: Theory and experimental evaluation,” Mathematical Programming, vol. 73, no. 2, 1996, pp. 129–174.

[15] P. Biswas, P. K. Mishra, and N. C. Mahanti, “Computational Efficiency of Optimized Shortest Path Algorithms,” International Journal of Computer Science & Applications, vol. 2, no. 2, 2005, pp. 22–37.

[16] C. Demetrescu, S. Emiliozzi, and G. F. Italiano, “Experimental Analysis of Dynamic All Pairs Shortest Path Algorithms,” ACM Transactions on Algorithms, vol. 2, no. 4, 2006, pp. 578–601.

[17] R. K. Ahuja, T. L. Magnati, and J. B. Orlin, Network Flows: Theory, Algorithms and Applications. Englewood Cliffs, N.J.: Prentice-Hall, 1993.

[18] K. Stachowiak, J. Weissenberg, and P. Zwierzykowski, “Lagrangian relaxation in the multicriterial routing,” in IEEE AFRICON, Livingstone, Zambia, September 2011, pp. 1–6.

[19] B. Musznicki, M. Tomczak, and P. Zwierzykowski, “Dijkstra-based Localized Multicast Routing in Wireless Sensor Networks,” in Proceedings of CSNDSP 2012, 8th IEEE, IET International Symposium on Communication Systems, Networks and Digital Signal Processing, Poznań, Poland, 18–20 July 2012.

[20] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, Introduction to Algorithms. MIT Press, 1990.

[21] D. P. Bertsekas, Network Optimization: Continuos and Discrete Models. Belmont, Massechusetts: Athena Scientific, 1998.

[22] E. W. Dijkstra, “A note on two problems in connexion with graphs,” Numerische Mathematik, vol. 1, 1959, pp. 269–271.

[23] S. Saunders, “A Comparison of Data Structures for Dijkstra’s Single Source Shortest Path Algorithm,” 5 November 1999.

[24] R. B. Dial, “Algorithm 360: shortest-path forest with topological ordering,” Communications of the ACM, vol. 12, November 1969, pp. 632–633.

[25] J. Bang-Jensen and G. Gutin, Digraphs: Theory, Algorithms and Applications. London: Springer-Verlag, December 2008.

[26] U. Pape, “Implementation and efficiency of Moore-algorithms for the shortest route problem,” Mathematical Programming, vol. 7, no. 1, 1974, pp. 212–222.

[27] D. P. Bertsekas, “A Simple and Fast Label Correcting Algorithm for Shortest Paths,” Networks, vol. 23, 1993, pp. 703–709.

[28] D. P. Bertsekas, F. Guerriero, and R. Musmanno, “Parallel asynchronous labelcorrecting methods for shortest paths,” Journal of Optimization Theory and Applications, vol. 88, February 1996, pp. 297–320.

[29] E. Dekel, D. Nassimi, and S. Sahni, “Parallel Matrix and Graph Algorithms,” SIAM Journal on Computing, vol. 10, no. 4, 1981, pp. 657–675.

[30] R. W. Floyd, “Algorithm 97: Shortest path,” Communications of the ACM, vol. 5, June 1962, p. 345.

[31] S. Warshall, “A Theorem on Boolean Matrices,” Journal of the ACM, vol. 9, January 1962, pp. 11–12.

[32] D. B. Johnson, “Efficient Algorithms for Shortest Paths in Sparse Networks,” Journal of the ACM, vol. 24, January 1977, pp. 1–13.

 ISBN: 978-1-61208-279-0

Handling Topology Updates in a Dynamic Tool for Support of Bandwidth on Demand Service

Christos Bouras, Ioannis Kalligeros and Kostas Stamos
Computer Technology Institute and Press "Diophantus"
Computer Engineering and Informatics Department, University of Patras
Patra, Greece
bouras@cti.gr, kallige@ceid.upatras.gr, stamos@cti.gr

***Abstract*—Automated Bandwidth Allocation across Heterogeneous Networks (AutoBAHN) is a tool under active development that supports a Bandwidth on Demand (BoD) service, intended to operate in a multi-domain environment using heterogeneous transmission technologies. The AutoBAHN system aims at providing a guaranteed capacity, connection-oriented service between two end points. Due to the distributed nature of the system and the fact that AutoBAHN has access to critical parts of the network, the importance of a robust, secure and dynamic mechanism for handling and distributing topology information cannot be underestimated. We will present how AutoBAHN manages to create an abstract representation of the physical network topology by hiding information such as IP addresses, port values, VLAN (Virtual Local Area Network), MTU (Maximum Transmission Unit), link capacities etc., how this abstract topology is merged among domains and how AutoBAHN system handles topology changes in terms of their influence on reservation processing.**

Keywords-Bandwidth on Demand; Topology Update; Topology Abstarction; Lookup Service; Quality of Service

I. INTRODUCTION

The GN3 European project is a research project funded by the European Union and Europe's National Research and Education Networks (NRENs). It is a continuation of the previous GN2 project and aims at building and supporting the next generation of the pan-European research and education network, which connects universities, institutions and other research and educational organizations around Europe and interconnects them to the rest of the Internet using high-speed backbone connections.

In the context of this project, a BoD service is being developed and the service is supported by the AutoBAHN tool. This BoD service is an end-to-end, point-to-point bidirectional connectivity service for data transport. It allows users to reserve bandwidth on demand between the participating end points. This service is offered collaboratively by GÉANT and a set of adjacent domains (NRENs or external partners) that adhere to the requirements of the service. These joint networks form a multi-domain area where the service is provided. The service offers a high security level in the sense that the carried traffic is isolated at the logical layer from other traffic.

The AutoBAHN system is capable of provisioning circuits in heterogeneous, multi-domain environments that constitute the European academic and research space and allows for both immediate and advanced circuit reservations. The overall architecture of the AutoBAHN system, its goal and the network mechanisms it employs are thoroughly presented in . AutoBAHN is responsible of managing the network elements in order to fulfill a user request. As a result, it needs to provide a mechanism for applying reservation actions on the necessary topology elements end-to-end. This paper highlights the structure of this mechanism and how AutoBAHN achieves, in spite its distributed nature, to have a total knowledge of the topologies of all NRENs that participate in this BoD service.

The rest of the paper is structured as follows: Section 2 presents the general architecture of the AutoBAHN system. In Section 3, we analyze how the topology handling mechanism works, while Section 4 presents in more detail how AutoBAHN processes topology changes and how they are propagated among the AutoBAHN instances. Sections 5 and 6 conclude the paper and present future fields of study.

II. AUTOBAHN ARCHITECTURE

The AutoBAHN system contains the Inter-Domain Manager (IDM), a module responsible for inter-domain operations of circuit reservations on behalf of a domain. This includes inter-domain communication, resource negotiations with adjacent domains, request handling and topology advertisements.

Furthermore, in order to build a real end-to-end circuit, the Domain Manager (DM) is another module that manages intra-domain resources. The IDM has an interface to the local DM (Idm2Dm), which undertakes all intra-domain functions (abstraction of the topology towards the IDM, scheduling and pre-reserving resources, monitoring etc.). This southbound interface of the IDM is the part of the AutoBAHN system that needs to be tailored to the domain-specific conditions.

In each domain, the data plane is controlled by the DM module using a range of techniques, including interfaces to the Network Management System (NMS), signaling protocols or direct communication to network elements. As part of AutoBAHN, a dedicated and independent Technology Proxy (TP) module allows the support of a range of technologies and vendors according to domain and global requirements.

 ISBN: 978-1-61208-279-0

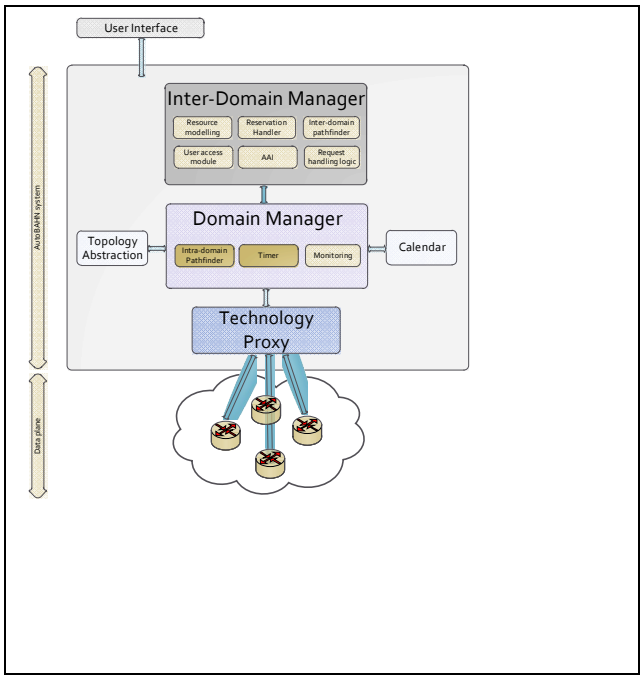


Figure 1. Basic architecture of AutoBAHN

The local NMS or service provisioning system, monitoring infrastructure, administration policies and security, may need to be adjusted for each networking domain making each TP implementation and configuration unique. However, the design of the DM has been optimized to support modular deployment and leverage the management infrastructure already deployed in any domain.

The above set of modules is deployed in each domain (NREN) that participates in the BoD service. A web based graphical environment (WebGUI) is used as a centralized portal for user access to the whole set of deployed instances.

III. Topology Handling

A. Storage

AutoBAHN uses cNIS , which is another service in the context of GN3 project in order to store and retrieve network topology elements.

The aim of the cNIS is to provide a unified repository for all relevant network information of a single administrative domain. cNIS was expected to be the "single point of storage", but in fact it is more than just a database. Apart from the internal functionality required for populating, validating and updating the database, it is equipped with modules for analyzing the topology data and presenting the data in a client-specified format (graphical, tabular or even XML (eXtensible Markup Language) for external applications). It can be used either as a component to build higher-level services and applications, or as a standalone repository of network topology data supporting network engineers in their daily administration work.

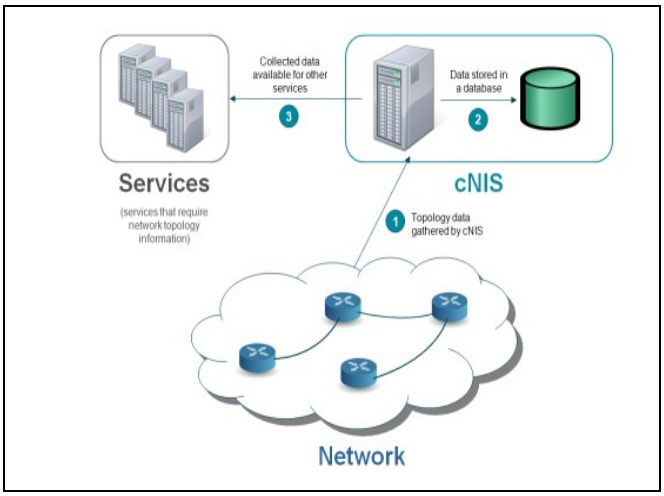


Figure 2. Basic usage of cNIS

As presented in Figure 2, cNIS implements three high-level functions, each covered by specific tools:

- Network topology data collection (automated or manual).
- Topology data management (keeping data consistent and up-to-date).
- Data sharing (interacting with external applications).

AutoBAHN retrieves the topology using a dedicated web service interface, stores it in its local database and creates an abstract representation which is later sent to Lookup Service (LS). LS is a common registry space for all AutoBAHN instances in order to exchange topology information. This procedure is thoroughly presented below.

B. Topology Abstraction

As it is already mentioned, topology elements and network resources are parts of the AutoBAHN system, dealing with reservation requests. There are three types of topology information:

- Physical network topology outside of the AutoBAHN system which is stored in cNIS.
- Technology-specific topology (DM level)
- Abstract topology (IDM level)

Due to the nature of the BoD service that AutoBAHN has to support, which is comprised of multiple heterogeneous and administratively separate domains, it is not possible, both for practical and for policy reasons, to handle global service topology in a flat manner and share topology information across the service. For example, one domain's administrators may not be willing to share details about their topology to everyone else participating in the BoD service. Therefore, detailed and technology-specific topology information is only kept per-domain, at the local cNIS instance. This topology is then abstracted before shared with the rest of the domains participating in the BoD service. This abstraction also enhances the heterogeneity of

 ISBN: 978-1-61208-279-0

the service since different network technologies can be represented in a unified manner and exchange information about the elements without taking into consideration how each network is implemented. Some of the supporting network technologies are Ethernet and SDH (Synchronous Digital Hierarchy) with the ability to easily extend current implementations to support also other technologies. The abstracted topology contains no technology-specific information such as actual interface names, VLANs, MTU sizes etc. It looks like a generic graph, uses ad hoc identifiers for network elements and it is shared among all domains. Each domain merges its own abstracted topology with the abstracted topology views advertised by the other domains, and this procedure converges with all domains having a shared view of the global abstracted topology.

The abstraction process that may be used by AutoBAHN is generally replaceable. The current abstraction algorithm creates one node for each edge node in the actual topology and then builds a virtual mesh network between them. The reason for choosing this algorithm is that it matches closely the 2-step way that AutoBAHN processes reservation requests: it deals with inter-domain links and domain edge nodes during the initial inter-domain pathfinding, and later each domain is responsible for finding a suitable intra-domain path. Inter-domain pathfinding operates on the globally shared abstracted topology, whereas intra-domain pathfinding is done separately by each domain and operates on the fully detailed local topology information.

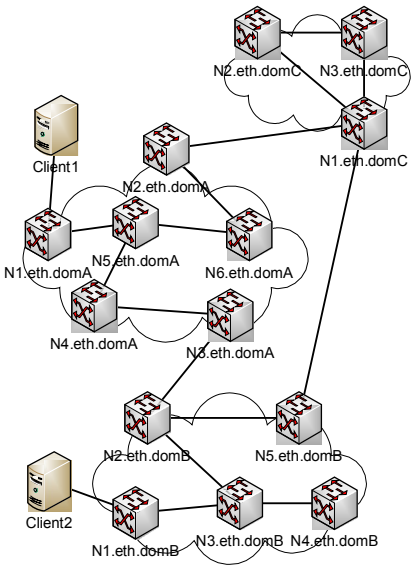


Figure 3. Technology-specific labels for topology components

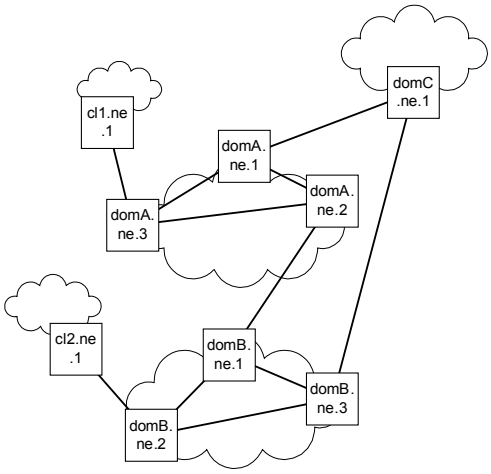


Figure 4. Abstracted topology view with global scope identifiers

Figure 3 shows a sample topology with domain-scope identifiers assigned to nodes and ports. The topology also contains information about interfaces, technical data including the half/full duplex attribute, MTU, capacity, and so on. Each domain only has local view of the technology-specific topology in Figure 3, for example domain C only knows about elements N1.eth.domC, N2.eth.domC, N3.eth.domC and their connections. The abstracted topology (that corresponds to the actual topology shown in Figure 3) is shown in Figure 4.

All domains have full knowledge of the global abstracted topology, so each domain is aware of Figure 4 in its entirety.

Abstract topology consists of domains, nodes, ports and links. Abstract identifiers for these elements are generated based on a 1-way hash of actual network elements in cases where corresponding actual network elements exist (for example for edge nodes) or artificial names in cases where corresponding actual network elements may not exist (for example for virtual mesh links in the abstracted topology).

The flow of the topology information within a single domain managed by AutoBAHN is the following; at the very beginning the domain's physical topology must be inserted into the technology-specific topology part by a system administrator. This topology includes a very detailed view of nodes, ports and links, also related to the resource usage and reservation mechanism. The DM operations are based on this technology specific topology view. However, as the IDM should not be aware of the technical details of the intra-domain topology and also the time issues within the domain, it must be provided with a less detailed, public

 ISBN: 978-1-61208-279-0

view, which can be also distributed to other domains. Once the DM has prepared the inter-domain abstract topology view, the latter is forwarded to the IDM module. It may be the whole topology during the system initialization phase, or it may be in the form of topology updates in the case of topology changes. In the second case, the IDM is responsible to forward changes that may affect one or more reservations. If a path is affected by a topology change (for example, by network failure), the reservation must be redirected to another path or must be aborted, depending on the resilience parameter specified by the reservation owner.

The IDM forwards the abstract topology view to the Lookup Service in order to have it advertised to AutoBAHN systems in neighboring domains. The advertisement mechanism also updates the local abstract inter-domain topology view. It contains not only the local view of the domain, but the entire inter-domain topology of all domains operating under a system, as it is retrieved by the LS. The abstract topology part of the system is the main source of information for the inter-domain pathfinding process.

C. Distribution

Since AutoBAHN is considered a distributed system because multiple instances are deployed across Europe, there is a need for those instances to be aware of the total network topology and not only the topology of the NREN that they actually belong. In order to fulfill this need, it was decided to have a centralized point of reference for all instances that could be used to exchange topology information and advertise their presence to other instances.

This is done with the utilization of Lookup Service which is also part of another GN3 project and more specifically perfSONAR . The Lookup Service acts as a service directory, where services (in our case AutoBAHN instances) can advertise themselves (provide their lookup information) and requestors are able to find any service they need. A more detailed view of Lookup Service's architecture and its purpose is described in .

Whenever an AutoBAHN instance starts, first of all, it writes an entry in LS with the following main attributes;

- Domain Name: unique identifier, e.g., GRNET, GEANT, PIONIER etc.
- URL: This is the web service endpoint of the domain's IDM interface through which other instances can use in order to communicate with one another

The second step is to query the LS to find what other instances have created a record. In that way, AutoBAHN knows what other instances are operating in the "neighborhood", stores this information in the local database and then communicates directly with them (if it is needed) through the interdomain (Idm2Idm) interface.

The procedure of adding new topology to LS is the following:

1. Check if LS contains existing abstract topology:

a. If yes, fetch it from LS and merge the elements that belong to other domains with the local elements. The final result first is stored to the local database and then is being written from scratch to LS.

b. If no, then the local domain's abstract topology is being sent to the LS for other domains to fetch it.

2. Update the timestamp of the entry.

IV. Topology Updates

As was described in the previous section, AutoBAHN creates an abstract representation of the topology. This is done by applying specific rules regarding the kind of the network element and a unique identifier that is created by information of this element such as IP address, name, domain where it belongs etc. The benefit of generating the identifier this way is that identifier generation is deterministic and constant among subsequent executions of the abstraction algorithm, even if the topology has changed significantly. In other words, the abstract identifier is kind of a signature for the network element that remains constant as long as the element is present in the topology.

For example, suppose that we have a network interface (port) with ID ath2.1-multi-10gbps that belongs to GRNET domain and a client (in other words an endpoint) that connects directly to this port. The hashing algorithm based on the id creates an 8 character long representation of this element which constitutes its abstract form. The final identifier becomes GRNET.pc.h438jd7t. The middle part is a categorization flag. In the specific example it means Port Client. AutoBAHN utilizes similar flags for other types of elements such as Port Interdomain (pi), Port Virtual (pv), Node Edge (ne) etc. In that way we have a quick solution for retrieving specific type elements from the abstract topology to process them. Furthermore, since this is a deterministic procedure, each time the algorithm is applied, the result is the same. This makes handling of topology updates and historical tracking of elements much simpler.was decided in order not to have database relations between actual network topology elements and abstract identifiers which would result in difficult handling in cases where after a topology update some elements could be removed by the topology. The latter is described in more detail below.

It is expected that during a topology update, AutoBAHN will have some reservations, either in an ongoing processing stage or at a processed (and finished) stage. This posed the problem of how to properly reflect the evolution of the topology, as network elements that may have been used or may be actively used by a reservation are modified or completely removed from subsequent instances of the domain topology. In order to handle this, we have chosen a different approach depending on whether a reservation is still being processed or whether it has finished processing.

Reservations that have been processed and they have terminated in some permanent state (e.g., finished

 ISBN: 978-1-61208-279-0

successfully, failed permanently or were canceled) are transferred to "historical" status. Historical information is persisted in the AutoBAHN database separately from present topology and reservation information, and is therefore immune to any topology changes.

Reservations that are currently being processed may be affected by the topology update, if for example their chosen path is no longer available, or if a better option has been made available. In order to handle these cases, AutoBAHN uses the following approach:

1. Check if new network topology is available in cNIS.
2. If yes, freeze currently processed reservations (with all their state) in a different persistence location.
3. Clean the existing network topology information.
4. Import new topology information from cNIS.
5. Abstract the new topology.
6. Share and merge the updated abstract topology with all other domains through the LS.
7. Restore frozen reservations and try to map them to the new topology.

Restoration of frozen reservations differs depending on the way that they are affected by the topology change, Reservations that depended on elements that were removed are currently rejected, although it is planned to implement a restoration mechanism in this case. Also reservations that are not directly affected are restored and continue being processed without any change, even if the updated topology presents opportunities for different path selection.

V. Conclusion and Future Work

AutoBAHN is being developed towards the goal of a formidable tool for supporting a bandwidth on demand service. A very important part of its success depends on its capability to dynamically adapt to changes and in particular to topology updates, which are in practice frequent and sometimes wide-ranging, without negatively affecting service availability. In this paper we have discussed how this goal has been pursued until now.

Future work includes the design and implementation of the capability to restore reservations that are affected by a topology update and their admission decision is no longer valid (if for example some of the elements on their previously selected path are no longer available). Such restoration could be to recalculate a new path from scratch based on the updated topology. Right now AutoBAHN utilizes a fallback mechanism. If the topology update results in a reservation processing failure, it provides the administrator the ability to restore AutoBAHN in its previous state right before the update. So, the administrator can take proper steps to handle this case in a manual manner, e.g., to cancel existing active reservations from the system that will conclude in failure after the update.

References

[1] "GN3 European Project," [Online]. Available: http://www.geant.net/pages/home.aspx. [retrieved: January, 2013]

[2] Mauro Campanella, Radek Krzywania, Victor Reijs, Afrodite Sevasti, Kostas Stamos, Chrysostomos Tziouvaras, and Dave Wilson, "Bandwidth on Demand Services for European Research and Education Networks," in 1st IEEE International Workshop on Bandwidth on Demand, San Francisco (USA), 2006, pp. 65-72.

[3] "GN2 Project", [Online]. Available: http://www.geant2.net/ [retrieved: January, 2013]

[4] "TERENA", [Online]. Available: http://www.terena.org/ [retrieved: January, 2013]

[5] "cNIS (Common Network Information Service)", [Online]. Available: http://www.geant.net/service/cnis/pages/home.aspx [retrieved: January, 2013]

[6] "perfSONAR", [Online]. Available: http://www.perfsonar.net/ [retrieved: January, 2013]

[7] Jason Zurawski, Jeff Boote, Eric Boyd, Maciej Glowiak, Andreas Hanemann, Martin Swany, and Szymon Trocha, "Hierarchically Federated Registration and Lookup within the perfSONAR Framework", in Proceedings of the 2007 Integrated Management Symposium, 2007, IFIP/IEEE, Munchen, Germany, May, 2007

 ISBN: 978-1-61208-279-0

Signature Generation Based on Executable Parts in Suspicious Packets

Daewon Kim, Jeongnyeo Kim, and Hyunsook Cho
Cyber Convergence Security Research Department
Electronics and Telecommunications Research Institute
Daejeon, Korea
{dwkim77, jnkim, hscho}@etri.re.kr

***Abstract*—Generally, attackers obtain the control authority of a remote host through the exploit/worm codes with some executable parts. The majority of the codes are still made of the codes which can be executed directly by CPU of the remote host without some decryptions. We focused on the fact that some parts in the exploit/worm codes include the function call related instruction patterns. In some suspicious packets with the exploit/worm codes, the function call instruction parts can be important information to generate the signature of Intrusion Detection System (IDS)/Intrusion Prevention System (IPS) for blocking the packets with the exploit/worm. In this paper, we propose the approach that detects the instruction patterns following the function call mechanism in some suspicious packets and generates a signature including the specific payload positions within the pattern-detected packets. We have implemented a prototype and evaluated it against a variety of the executable and non-executable codes. The results show that the proposed approach properly classifies the executable and non-executable codes and can generate the high-qualified signature based on the analyzed results.**

Keywords-network security; intrusion detection system; intrusion prevention system; malicious code; exploit code; worm code

I. INTRODUCTION

To avoid the signature-based IDS/IPS such as Snort [4], Bro [5] and recent techniques [11], [12], encrypted exploit/worm codes [1]-[3] are gradually increasing. However, in real fields, most of the exploit/worm codes are still non-encrypted codes. Therefore, it is possible to detect and prevent the exploit/worm codes if a distinction can be made between the executable and non-executable codes in network flows with the anomalous and suspicious traffic patterns because normal network services of servers are primarily based on non-executable plain texts and not executable codes [6],[7].

Several researches were published to detect malicious codes in network traffic. Earlybird [8] and Autograph [9] are based on the fact that different instances of the exploit/worm codes would contain common substrings or fingerprints, which would potentially have the code patterns to penetrate vulnerabilities. TRW (Threshold Random Walk) [10] is based on the idea that the exploit/worm codes infected host that is scanning the network randomly will have a higher connection failure rate than a host engaged in legitimate operations. However, for generating signatures, the above researches have difficulty analyzing the logical features of non-encrypted malicious codes because they are based on the simple matching of repeated payload substring and traffic-behavior. As a result, the probability of detection decreases significantly as the size of input data is decreased.

Although not a complete program, the executable part of a non-encrypted malicious code has very logical features. As a malicious code has many action roles, attackers have included many function-based logics in malicious codes. Finally, non-encrypted malicious codes have high probability of including the logical feature following function call mechanisms.

In this paper, we extended our previous work [13] by proposing a signature generation method based on the payload positions detected by our function call detection mechanism. The proposed method calculates the match probabilities of instruction patterns according to the function call mechanism and determines the existence of executable codes in the suspicious packets of anomalous traffic. Finally, the method generates a unique signature with the packet payload including the detected function call instructions.

The rest of the paper is organized as follows: Section II overviews the background and operation according to our method; Section III presents analysis steps of the proposed method; Section IV shows the experimental results; and Section V presents our conclusion and suggestions for future

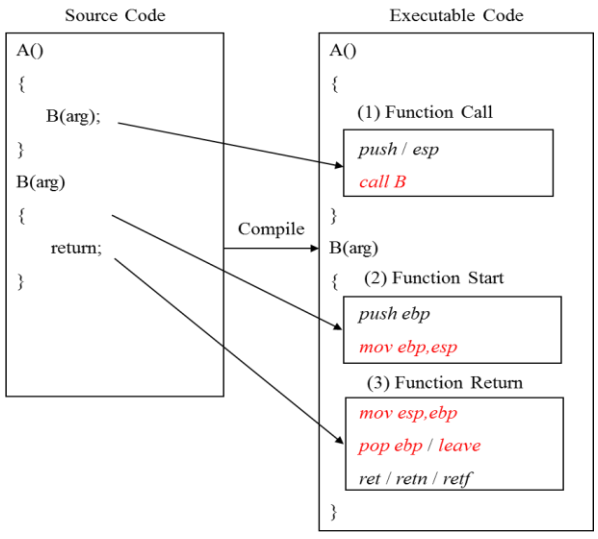


Figure 1. Instruction patterns of function call/return pairs.

 ISBN: 978-1-61208-279-0

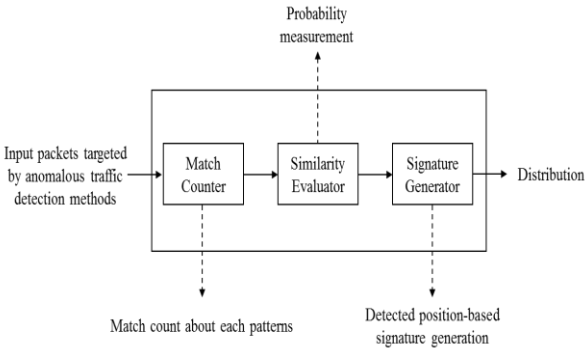


Figure 2. Operational overview.

TABLE I. INSTRUCTION PATTERNS ACCORDING TO FUNCTION CALL MECHANISM

Fn.	Not.	Instruction order to be matched for each notation		
		1	2	3
Call	ec	esp ops.	call(s)	
	pc	push	call(s)	
Start	pm	push ebp	mov ebp,esp	
Return	mpr	mov esp,ebp	pop ebp	ret(s)
	pr	pop ebp	ret(s)	
	lr	leave	ret(s)	

'esp ops.' means instructions that include '*%esp*'.

works.

II. OVERVIEW

The function call instruction patterns are one of the logical features in the executable codes. If a source code with functions is compiled, the function parts are transformed into the instruction patterns with *call*/*return* pairs. In the IA (Intel Architecture)-32, Fig. 1 shows the generated instruction patterns after the function *call*/*return* is compiled.

The proposed method detects the patterns of Fig. 1 and decides in terms of probability whether an executable code exists in the payload of suspicious packets or not. After that, the method generates a signature based on the detected position in the packets. Fig. 2 shows the simple process flows of the method.

In Fig. 2, *Match Counter* measures the trial and match counts of Fig. 1 instructions about the input packets. *Similarity Evaluator* has the pattern match probabilities of executable codes and compares them with the results of *Match Counter*. *Signature Generator* generates a signature including the payload around detected positions.

III. SIGNATURE GENERATION BASED ON FUNCTION CALL INSTRUCTIONS

A. Match Counter

The pattern match counts in the detection window of any instruction range are measured, and moving through the instructions one by one, this measuring is repeated to the end

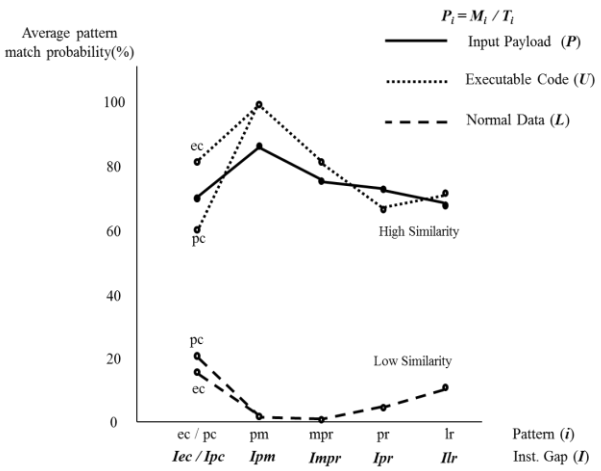


Figure 3. Probability comparison of input payload and real executable code.

of an input payload. In the IA-32, the instruction patterns defined by our method are presented in Table I.

In Table I, the attempt to match patterns is triggered by the gray-highlighted triggering instructions. Other instructions in Table I are inappropriate for triggering detection because they are frequently appeared regardless of the function call mechanism. When *Match Counter* measures the trial and match counts according to the instruction patterns of Table I, the instruction gaps between the instructions of Table I have to be considered. It is because some additional instructions can be made between the instructions of Table I by a compiler. Therefore, the pattern match counts of Table I should be counted in the acceptable instruction gap size.

In the case of a *pc* pattern, for example, if the *call* - which is the instruction number 2 - within the detection window is detected, the trial count of the *pc* pattern is increased by one. If the instructions are compared one by one in the reverse direction of the *call*, and if the *push* - which is the instruction number 1 - is detected, the match count is increased by one. At this time, if the number of instructions tracked as the reverse direction exceeds the pre-defined instruction gap size, the match count is not increased because of the match failure.

B. Similarity Evaluator

After the trial count set T and match count set M are measured on each notation, the match probability set $P_i = M_i / T_i$, where $i = \{ec, pc, pm, mpr, pr, lr\}$, is calculated. Our basic idea considers that P will be similar to the match probability set U of the real executable code if the input packets have executable codes constructed as some functions. Fig. 3 shows an example to describe this idea.

In Fig. 3, the match probability exists in both the executable and non-executable code. It means that the false positive can be large if the total trial count is very small. Therefore, for a more reliable analysis result, the similarity calculation to decide the existence of executable codes in the current detection window should be processed when the total trial

 ISBN: 978-1-61208-279-0

count within the current detection window is larger than the minimum trial count e.

The more similar the input payloads are to the executable code, the closer P would be to U. This could be calculated from the relative similarity set R_i between U and L like the below formula.

$$R_i = \frac{P_i - L_i}{U_i - L_i} \; except \; for \; T_i = zero, \quad (If \; P_i \geq U_i, \; R_i = 1.0. \; If \; P_i \leq L_i, \; R_i = zero.) \tag{1}$$

In Fig. 3, each pattern has individual probability gaps between U and L. It means that the decision about the input payload is more reliable if the gap is large. Therefore, the weight set W_i is required to increase the reliability.

$$W_i = \frac{U_i - L_i}{\sum_j U_j - L_j} \; except \; for \; T_i = zero, \tag{2}$$

If the final weighted similarity s is larger than the decision threshold d, the input payload evaluated by our method has high probability of including some executable codes.

$$s = \sum_i (W_i \cdot R_i) \; except \; for \; T_i = zero. \tag{3}$$

C. Signature Generator

Signature generating does not require special techniques in this paper because the signature style is various according to the IDS/IPS. Based on the detection results of our method, it can be the entire payload or the specific-range payload in an input packet. In the case of specific-range, the signature needs to be a continuous range to include the detected all position.

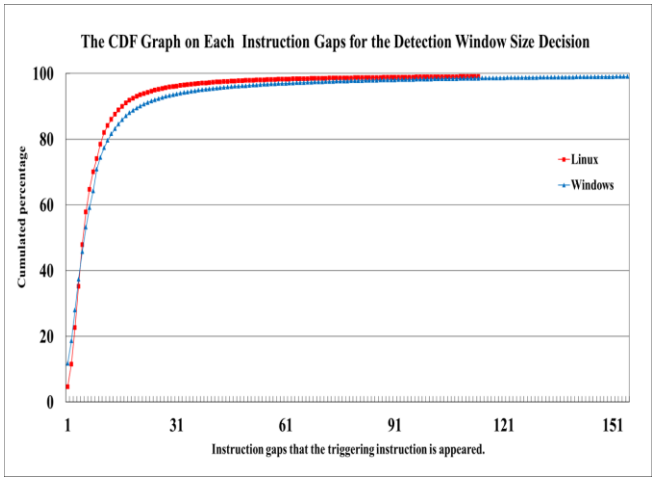


Figure 4. Existence probability of instructions gaps between triggering instructions.

IV. EXPERIMENTS

The test files for the experiments are the IA-32 based 3000 executable files of Windows/Linux and 3000 data files such as .txt, .doc, .ppt, pdf, mp3, gif, etc. In the case of executable files, only <.text> section was used in the experiment.

A. Size of Detection Window

When the detection window moves one byte at a time, the triggering instruction is always required for the analysis. In Fig. 4, when we select the existence probability of triggering instruction as 99%, the detection window sizes were 114 instructions in Windows and 155 in Linux. Therefore, the desired size *z* can be set as 155, which is about 450 bytes.

B. The Match Probabilities and Instruction Gaps of Executable and Non-Executable Code

Table II shows the experiment results for the match probabilities and the instruction gaps of executable and non-executable code. The determined detection window size of 450 bytes and the results of Table II show that this work proposes a reasonable method for detecting executable codes although the input is only one packet.

C. Executable Threshold and Minimum Trial Count

Figs. 5 and 6 show some parts of experimental results to determine the executable threshold d and minimum trial count e. In Fig. 5, the threshold d of executable codes is

TABLE II. PARAMETERS DETERMINED BY EXPERIMENTS

Notation	I	U	L
ec	1	0.80	0.10
pc	2	0.60	0.20
pm	3	0.98	0.02
mpr	2	0.80	0.01
pr	7	0.75	0.25
lr	2	0.70	0.10

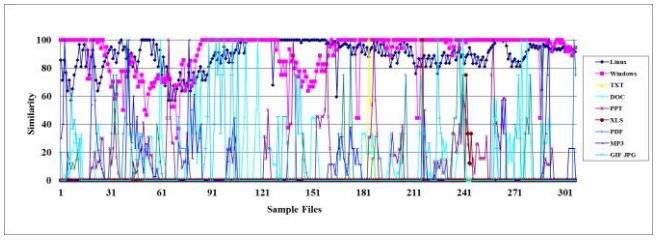


Figure 5. A graph for determining executable thresholds.

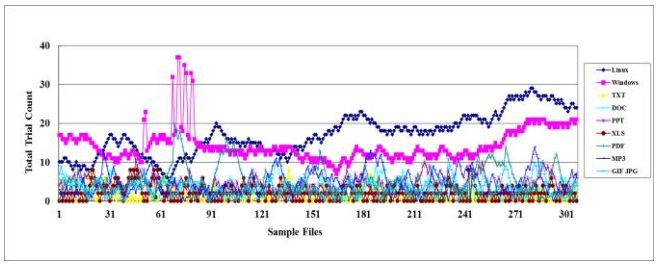


Figure 6. A graph for determining a minimum trial count.

 ISBN: 978-1-61208-279-0

over 60% and in Fig. 6, the minimum trial count e is about 3.

V. CONCLUSION AND FUTURE WORKS

In this paper, we proposed an approach that detects the instruction patterns following the function call mechanism in some suspicious packets and generates a signature including the specific payload positions within the pattern-detected packets. As the experiments shows, the proposed detection method is efficient even for one packet.

Regarding the method of detecting executable codes, our method analyzes in a form that is similar to the pattern-matching of instruction patterns following the function call mechanism. Our method can determine whether the executable codes exist or not in terms of the probability even in small input payload. In current method, we used a Detection Window of several hundred bytes. In next experiment, we will try a method which sequentially searches the payloads in order to detect the triggering instructions without the Detection Window. As a result, we may be able to identify the function call patterns for input payloads of a smaller size.

ACKNOWLEDGMENT

This work was supported by the ETRI R&D program of KCC (Korea Communications Commission), Korea [12-912-01-001, "Development of MTM-based Security Core Technology for Prevention of Information Leakage in Smart Devices"].

REFERENCES

[1] M. Polychronakis, E. P. Markatos, and K. G. Anagnostakis, "Network-level polymorphic shellcode detection using emulation," Proc. of the Third Conference on Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA06), July 2006, pp. 54–73.

[2] Q. Zhang, D. S. Reeves, P. Ning, and S. P. Lyer, "Analyzing network traffic to detect self-decrypting exploit code," Proc. of the ACM Symposium on Information, Computer and Communications Security (ASIACCS07), 2007, pp. 4-12.

[3] M. Polychronakis, E. P. Markatos, and K. G. Anagnostakis, "Emulation-based Detection of Non self-contained Polymorphic Shellcode," Proc. of the International Symposium on Recent Advanced in Intrusion Detection (RAID07), 2007, pp. 87-106.

[4] M. Roesch, "Snort: Lightweight intrusion detection for networks.," USENIX LISA Conference, 1999, pp. 229-238.

[5] V. Paxson, "Bro: a System for Detecting Network Intruders in Real-time," Proc. of the USENIX Security Symposium, Jan. 1998, pp. 2435-2463.

[6] R. Chinchani and E. V. D. Berg, "A fast static analysis approach to detect exploit code inside network flows," Proc. of 8th International Symposium on Recent Advanced in Intrusion Detection (RAID05), 2005, pp. 284-308.

[7] X. Wang, C. Pan, P. Liu, and S. Zhu, "SigFree: A Signature-free Buffer Overflow Attack Blocker," Proc. of the 15th USENIX Security Symposium, 2006, pp. 225-240.

[8] S. Singh, C. Estan, G. Varghese, and S. Savage, "Automated worm fingerprinting," Proc. of the 6th Symposium on Operating Systems Design & Implementation (OSDI04), 2004, pp. 45-60.

[9] H.-A. Kim and B. Karp, "Autograph: Toward automated, distributed worm signature detection," Proc. of the 13th USENIX Security Symposium, 2004, pp. 271–286.

[10] J. Jung, V. Paxson, A. W. Berger, and H. Balakrishnan, "Fast portscan detection using sequential hypothesis testing," Proc. of IEEE Symposium on Security and Privacy, 2004, pp. 211-225.

[11] B.-H Chang and C. Jeong, "An Efficient Network Attack Visualization Using Security Quad and Cube," ETRI Journal, vol. 33, no. 5, Oct. 2011, pp. 770-779.

[12] S. A. Taghanaki, M. R. Ansari, B. Z. Dehkordi, and S. A. Mousavi, "Nonlinear Feature Transformation and Genetic Feature Selection: Improving System Security and Decreasing Computational Cost," ETRI Journal, vol. 34, No. 6, Dec. 2012, pp. 847-857.

[13] D. Kim, Y. Choi, I. Kim, J. Oh, and J. Jang, "Function Call Mechanism Based Executable Code Detection for the Network Security," Proc. of the International Symposium on Applications and the Internet (SAINT08), 2008, pp. 62-67.

 ISBN: 978-1-61208-279-0

A Distributed Power Management Algorithm for a Self-optimizing WiFi Network

Abheek Saha
Hughes Systique Corp.
Gurgaon, India
Email:abheek.saha@hsc.com

***Abstract*—In this paper, we propose a fully distributed algorithm for a dense unplanned self-optimizing network of 802.11 access points. As opposed to the traditional method of having a centralized controller to collect information from all access points, our algorithm runs independently on each AP to create a cooperative network of access points, with no explicit inter-node communication. We present simulation and actual laboratory data which shows how our algorithm can significantly improve network performance in a robust and scalable manner.**

***Keywords*—Self optimizing networks; distributed algorithms; cooperative games; femtocells**

I. INTRODUCTION

A self optimizing network (SON) of wireless nodes represents a collection or network of co-located network nodes which can jointly set their own operating configuration so as to maximize network performance using suitably chosen parameters. SON controllers for WiFi Access Points (APs) have been launched by Aruba [1], Ruckus [2] and others. The CISCO suite of Wireless LAN (WLAN) applications includes the Cisco Unified Wireless Network module, which incorporates various SON features. In the LTE domain, SON has been included into the 3gPP specifications [3] and is being actively pursued by Nokia Networks, Ericsson, AT&T and others.

In a typical self-optimizing network, we are interested in optimizing some global performance metric, subject to constraints in another global metric, by controlling specific operational parameters at each individual node in the system. For example, a SON can try to minimize handovers in the network as a whole, subject to average call drop probability being above a certain threshold; both of these are global, user-visible metrices. The optimization is achieved by setting certain policy parameters in each network node; for the above example, it may be the Signal and Interference to Noise Ratio (SINR) threshold at which a handover is triggered.

In the standard self-optimizing networks as discussed in [4]–[7], the SON function resides in a central entity (or cluster of entities) called the controller. The controller receives feedback from the network nodes (base-stations, access points, or node B, depending on the Radio access technology) and in turn computes and sends configurations to the network nodes; hence translating global network state information to local control values for each network node. We call this the *central control* model.

The object of this paper is to propose an alternate, distributed model for self-optimizing algorithms in modern wireless systems. Our work is on the same lines as [4] [8] but our model and technique vary significantly. We take a fully distributed approach to our problem by borrowing elements from the most successful distributed control applications that we are aware of. The result is a simple, scalable algorithm, which works in a multitude of conditions. We have simulated this and implemented it in real life using a network of WiFi access points, by using the basic interface provided by the Linux *hostapd* application. With the recent release of the Femto API by the small cell forum, we believe that SON applications can be implemented very easily on the network cell nodes themselves. A fully distributed algorithm is a significant step to solving the problems of connectivity, scalability and robustness which limits centralized algorithms.

The rest of this paper is organized as follows. In Section II, we describe the details of transmit power control in a WiFi network. In Section III, we describe the problem to be solved in the context of the network topology and network node properties, for which our algorithm is presented. In Section IV, we describe the basis for the distributed algorithm and its adaptation for our particular problem. In Section V, we describe the design of the test-bed and the validation of the algorithm. Finally, in Section VI, we conclude our analysis and areas of future work.

II. SETTING TRANSMIT POWER IN A WIFI NETWORK

While there are many wireless network management functions which are suitable for SON applications, a commonly used one is interference and transmit power management. This single function covers a great deal of ground; by adjusting beacon power, we can adapt coverage, by adjusting traffic channel power, we can adjust SINR and throughput and finally, by a combination of the two, we can also adjust handoff performance.

A. The impact of transmit power setting

What are the implications of adjusting transmit power? A network node broadly transmits three kinds of waveforms. First is the beacon and/or pilot channel. This transmission is necessary for mobile devices in the idle state to detect the presence the network node and associate with it - thus, the pilot/beacon power controls the coverage or reach of the signal. This signal is also used by user terminals during handover as a way of identifying the relative strength of a network node compared to the one it is currently *camped on*. Pilots are typically at significantly higher power than the rest of the transmission, but occupy a relatively small part of the overall

 ISBN: 978-1-61208-279-0

power budget, because they occupy a small portion of the available spectrum.

The second power level is that used for transmission on the traffic channels; this is spread over embedded pilot, data and control bits. Typically, the embedded pilot bits are transmitted at a certain level above/below that of the traffic bits , since they have to be received reliably in order to do accurate channel estimation. A power budgeting function is used to allocate energy per bit so as to maintain these relative ratios as per network configuration. Control bits are usually protected through coding, rather than power setting and thus are at the same power level as data bits. The quality of the service that a given user terminal receives is a function of the SINR on the data channels (subject to the signal level itself being above the receiver sensitivity threshold; but this condition is very mild thanks to the astonishing sensitivity of modern receivers). As the SINR rises, the individual bit error rate (BER) drops, and consequently the user terminal can increase the transmission rate (by reducing protection and/or using more aggressive modulation schemes); this is known as adaptive modulation and coding. It should be noted, however, that the relationship between SINR and transmission rate is not linear. For example, at a given coding rate and scheme, the block error rate (BLER) is only affected by the SINR once the number of bits in error per block cross a certain threshold value known as the *free distance* of the coding scheme. While there are many studies on this topic, a log-linear curve has been shown to hold true in a number of cases [9].

III. ALGORITHM CONTEXT

We place our algorithm in the fairly generic context of a network of 'network nodes', 'mobile/user devices' and an allocated set of frequencies, which are the shared physical resource. The purpose of the network is to allocate frequencies to individual nodes, so as to offer the maximum quantum of service (measured both in terms of throughput and QoS) to active users. The number of frequencies available is typically fixed, whereas the network can scale indefinitely, both in terms of network nodes and in terms of actual users. Gupta and Kumar point out in [10] that in this kind of capacity limited, interference constrained network, the aggregate throughput can at best scale at $\sqrt{n}$, whereas the throughput per user will actually scale at $\frac{1}{\sqrt{n}}$. We wish to find an algorithm to achieve this target.

A. Properties of network components

We use the term network node to describe eNodeB/Access Points. The common characteristics of these nodes are:

- They transmit a mix of pilot and data signals on the downlink.
- The total power required for data traffic is determined by the mix of users they are supporting and their distances from the node itself
- The network nodes use a fair allocation scheduler, so that the traffic transmission is not dominated by the best placed user devices, but the average.
- The network node is aware of its own position.
- The utility of the network node is determined by the average bit-rate it supports and the average number of active users.
- The network node has a transmit power level u_i which allocates the combined power for each resource it is using. It is power-limited in the sense that the value u_i must be less than some maximum power level P_u. This comparision may be done instantaneously, i.e., $u_i(t) \leq P_u \forall t$ or averaged over a slot or a frame. A power allocation of zero for a given network node indicates that the node is switched off.
- The network node has fine grained control over its transmit power level, i.e., it can adjust transmit power on a frame by frame basis [11] and allocate variable power to control and data frames; however, it has also been pointed out that this has limitations based on the technology and realization thereof [12]. The network node is free to manage the individual allocations to frames/users/channels as long as the overall power budget is maintained.

We further make the following reasonable assumptions about the topology

- The network nodes are randomly placed in a 'dense' environment, i.e., theoretically, any of the clients may attach to any of the nodes
- Each network node is assigned one channel on startup; more channels can be allocated depending on availability

Each network node has a set of users (WiFi clients) to whom they are offering data transport services. The user population is assumed to have the following general principles

- The user devices are located randomly in the area; for simplicity's sake a given user device can pick absolute any network node in this area, subject to operational constraints, i.e., sufficient RSSI and SINR
- The user devices are homogenous; they are characterized by a data generating process with common properties

We say that a network node is *backlogged* if at least one of its users is continuously backlogged, i.e., has a non-zero queue for the duration of the measurement T. In practical terms, if a network node has no backlogged users then it should reduce its traffic load, because it is possibly offering more service than its users are demanding. Given the set of backlogged users, we assume that the network computes a target SINR value τ_i which represents the ensemble of backlogged users. The utility for the network is given by the function

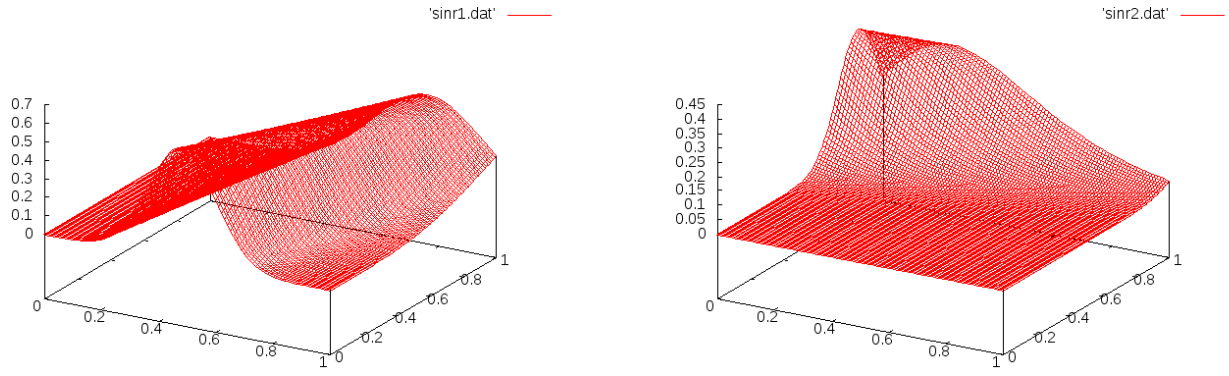
$$U(u_i) = \mathbb{E}_k e^{-\left[(s_{i,k}-\tau_{i,k})^2\right]} \quad (1)$$

As we can see in Fig. 1, the utility function for a given network node is quasi-concave.

Each network node transmits at a level $u_i(t)$, which is the control variable for our algorithm. The signal received by the kth user of the ith network node thus becomes

$$S_{k,i} = \frac{u_i F(d_{k,i,i})}{\sigma_i + \sum_{j \neq i} \gamma_{i,j} u_j F(d_{k,i,j})} \quad (2)$$

 ISBN: 978-1-61208-279-0



(a) AP 1 (b) AP 2

Figure 1. Utility functions for two access points

where $d_{k,i,j}$ is the distance between the kth user of the ith Access Point from the jth access point and $F()$ represents the fading. $\gamma_{a,b}$ is the interference coupling between access points a and b. If the two are using orthogonal frequencies, it is 0, if the two are using the same frequency, it becomes 1. For a given network node, we assume the existence of a metric function $G(i) = g(S_{1,i}, S_{2,i}, \ldots)$, which is monotonic in each of its elements. A simple metric function is the average or infimum.

B. Formulation of the standard problem

We start with the easiest form of the problem. Assume that $\alpha_{i,j}$ represents the marginal impact of increasing transmit power u_j for the ith Access Point. Therefore $\alpha_{i,j} = \mathbb{E}_k(\gamma_{i,j}F(d_{k,i,j}))$. The formulation in (2) can be rewritten as:

$$S_i = \mathbb{E}_k(S_i) = \frac{\alpha_{i,i}u_i}{n_i + \sum_{j\neq i}\alpha_{i,j}u_j} = u_i\kappa_i \quad (3)$$

For an equilibrum solution, we need to solve $\frac{\partial U_i(u_i)}{\partial u_i} = 0 \forall 0 \leq i < n$. To do this, we note that

$$\frac{\partial U_i}{\partial u_i} = \frac{\partial U_i}{\partial s_i}\frac{\partial s_i}{\partial u_i}$$

$$= -2.0(s_i - \tau_i)e^{-(s_i-\tau_i)^2}\frac{\alpha_{i,i}}{n_i + \sum_{j\neq i}\alpha_{i,j}u_i} \quad (4)$$

(5)

This has one solution where $s_i = \tau_i \forall i$. We can write the resultant set of equations as a matrix equation

$$\begin{bmatrix} u_0 \\ u_1 \\ \cdots \\ u_{n-1} \end{bmatrix} \begin{bmatrix} \alpha_{0,0} & -\tau_0\alpha_{0,1} & \cdots & -\tau_0\alpha_{0,n-1} \\ -\tau_1\alpha_{1,0} & \alpha_{1,1} & \cdots & -\tau_1\alpha_{1,n-1} \\ \cdots & \cdots & \cdots & \cdots \\ -\tau_n\alpha_{n-1,0} & \tau_{n-1}\alpha_{n-1,1} & \cdots & \alpha_{1,n-1} \end{bmatrix} = \begin{bmatrix} n_0\tau_0 \\ n_1\tau_1 \\ \cdots \\ n_{n-1}\tau_{n-1} \end{bmatrix}$$

The optimization problem as formulated in (6) is solvable under some basic conditions; specifically, the matrix has to be invertible. However, this solution requires the solver to have full state information of p_i and s_i for all network nodes. This can be acquired by querying individual network nodes, but requires a robust communication channel, scaling with the number of access points. A bigger challenge is to compute the values of $\alpha_{0\leq i<n, 0\leq j<n}$ and the values of the individual noise terms. In fact, there is no possible way to compute n, the noise terms and the values of $\alpha_{i,j}$, even under the fairly reasonable assumption that $\alpha_{i,j} = \alpha_{j,i}$, since we have a total of $2 * n$ readings and $n_{C_2} + n^2 + n$ unknowns. Most of the available solutions assume that the values of the noise term are known by external means [5].

Is it possible to compute the value of $\alpha_{i,j}$ using external means? Note that $\alpha_{i,j}$ consists of two terms, one being the attenuation caused by distance and the second being the coupling. For the latter we need to know the channels allocated to each node and how much ACI/CCI is being generated - this is relatively easy to get. However for the former, we need a channel model which captures the relationship between geographical position and the attenuation and other channel model parameters specific to that particular environment and geography. OFDM traffic models are complex and sensitive to environmental artefacts and this typically requires a lot of intervention from the user/administrator. In commercial models, the relative coupling between nodes are typically embodied in the ANR and are based on a combination of channel modeling and continuous measurements.

A second challenge is to compute a value of τ_i for each access point. Clearly, an excessively low value of τ will cause bad service, whereas an excessively high value of $\vec{\tau} = \{\tau_0, \tau_1, \ldots, \tau_{n-1}\}$ means that the matrix given in (6) is no longer positive definite, hence the program isn't solvable. In a centralized environment, we will be able to set $\vec{\tau}$ to the maximum set of values so as to allow a solution to (6). This is a semidefinite programming problem and is not very easy to solve either. Alternate game theoretic formulations have the

 ISBN: 978-1-61208-279-0

same results and similar issues.

We would thus like to avoid any solution which depends on static analysis or pre-configuration to compute the above values; ideally, our algorithm would be an iterative solution, capable of using empirical measurements to fine tune its estimate of the network state. If we can design an algorithm that only uses empirical measurements from individual network nodes, it lends itself to distributed implementation. In our experience, a distributed algorithm will typically score over any centralized solution in terms of scalability (the ability to handle larger and larger numbers of nodes), robustness and flexibility. The classic paper by Kleinrock and Tung [13] demonstrates a elegant and remarkable solution to what seems to be an intractable optimization problem. On the other hand, distributed algorithms need to be carefully designed so as to be stable and converge sufficiently quickly to the appropriate solution. Convergence time for different categories of distributed algorithms is an enormous research topic in its own right.

IV. A DISTRIBUTED ALGORITHM FOR POWER CONTROL

We propose a distributed power control algorithm loosely modeled on the well-studied TCP congestion control algorithm. There are some crucial similarities and some crucial differences, as we shall demonstrate in the subsequent sections.

A. TCP congestion control as a distributed optimization algorithm

The TCP congestion control algorithm is a extremely well studied and widely deployed example of network wide congestion control [14]. It has been extremely succesful in practice, and has benefited from many years of continuous refinement and innovation (TCP New Reno, TCP Vegas, TCP Westwood, Eifel and TCP CUBIC). Its strengths lie it its flexibility (starting from switched links of 56kb/s, it has been deployed in every concievable environment, including wireless, gigabit Ethernet, terabit optical fiber and satellite), its scalability (literally thousands of individual nodes and tens of thousands of connections) and its robustness. It is also easy to deploy, requiring practically no intervention or configuration from the user. Based on our own extensive experience of TCP, we find its success arises from three factors.

- It constantly switches between stability and network probing. This ensures that it never enters a sub-optimal local equilibrium and can also automatically adjust to changes in global conditions (network load, network bandwidth availability, etc.)
- It acts aggressively before it discovers congestion and conservatively on discovering congestion (fast attack/slow retreat). This allows new TCPs to enter and adjust rapidly to existent network conditions
- It constantly updates its estimates of the two critical parameters in the algorithm; the round trip latency (to avoid loop gain) and the congestion buffer size. There are no fixed or externally programmed thresholds.

TCP works by transmitting a number of packets into the network destined to a peer TCP and waiting for a set amount of time. If the packets are positively acknowledged within that time, it assumes that there is bandwidth/buffering available in the network and transmits a few more packets, this time a slightly higher number. If the pre-set time expires, it assumes that the transmitted packets were lost due to congestion in the network. So it reduces its transmission rate and tries resending the packets. Each TCP computes its own estimate of the network measures independently and applies its congestion algorithm independently of the others in the same shared network. The implementation is completely distributed; no TCP needs to exchange any information with any other TCP. Yet the system performs stably and scales, due to the features built into the algorithm itself.

B. A brief review of TCP

Leaving aside *congestion avoidance* for the moment, there are two variables which govern a particular TCP instance. The first is the transmission rate, measured by the number of packets to be injected into network per roundtrip time. TCPs call this variable *cwnd*; it is the internal estimate of the amount of buffering available in the network. The higher the *cwnd*, the faster the TCP transmits. The second is the estimate of roundtrip time, which is a statistical measure $\hat{rtt} = \mu_{rtt} + 2\sigma_{rtt}$, constantly updated by the measured timegap between a packet being transmitted and the acknowledgement being received for it. The variables represent the tension in the system; if the TCP underestimates the round trip time, it will time out prematurely and inject unnecessary packets into the network, adding to congestion. If the TCP over-estimates the round trip time, its reaction to network conditions and ability to recover from congestion is diminished.

There is a further purpose to the round trip measurement; it allows the TCP to estimate a *baseline* round trip time, below which the TCP cannot possibly expect feedback. This is the base latency of the network, the actual time taken for a packet to traverse the empty network. In addition, a component of the baseline rtt is the latency induced by background network traffic; network traffic which is independent of the actions of this TCP. TCP algorithms such as VEGAS [15] use variations of the measured round trip time, in conjunction to the knowledge about the baseline delay to detect congestion.

Further, the TCP congestion control algorithm uses *Additive Increase and Multiplicative Decrease*. Even though multiple TCPs may be transmitting in parallel, each TCP reacts to a congestion signal as if it was the sole contributor for this congestion and executes a multiplicative decrease of its traffic rate; this bypasses the distributed coordination issue. Empirical data shows that this conservatism is as the heart of TCP's success in maintaining network stability.

It is important to understand the parallels between TCP and our situation. As in TCP, we also have the loop time, which is basically the time taken for the rest of the network to detect and respond to any unilateral changes in power; its value depends on the frequency of measurement updates and the averaging period for the other Access Points in the network. The analogue to the congestion window is of course, the transmit power setting of the network node.

 ISBN: 978-1-61208-279-0

C. A local algorithm for power adjustment

The algorithm that we have devised is a mixture of the approaches suggested in the various TCP algorithms. As in NewReno, we use a congestion signal for backing off transmission power; in our case, the congestion event occurs when the measured noise is greater than the baseline noise by a fixed Δ amount. Like New Reno, we use 'bandwidth hunting', by constantly trying to adjust transmit power onwards. We have used some of the ideas in TCP CuBic to make the algorithm self timing.

There is an important issue to be solved, however, which is not addressed by New Reno. This is the problem of background noise - noise which is independent of the network nodes transmission levels or those of its neighbours. In our New Reno analogy, this is akin to having an additional amount of packet drops which operate independent of network node induced congestion. The assumption in New Reno is that all packet drops are caused by congestion - a valid assumption in normal wired networks, but one which causes some issues in high delay wireless networks [16].

In our case, baseline noise exists and it varies over time and networks. A network node, on entering the system, has to be able to detect baseline noise dynamically and adjust to it. An analogy exists in TCP Vegas, which works by mapping congestion to variations in round-trip delay. For this to work well, the Vegas algorithm needs to be able to estimate the base delay in the network. We adapt the same to our system, except we measure baseline noise. The measurement of baseline noise has an interesting ramification - specifically, the higher the estimate of baseline noise, the more aggressive an endpoint is going to be. This gives an additional incentive for network nodes to minimize their transmission power ; otherwise, network nodes which enter (or re-enter) the system are going to over-estimate the baseline noise initially and act more aggressively. This effect has been noted in Mo et al [17].

Secondly, there is no directly analogue of a utility function for a TCP connection - at least, one which is explicitly built into the algorithm. Rather a TCP connection attempts to maximize throughput, subject to a complex set of rules. However, there have been many studies of what utility function TCP is *effectively* using [18], [19]. It is obvious, that TCP pays more attention to the bandwidth delay product, than pure bandwidth; however, packet drops also impact it. In our case, we are attempting to maximize an explicit utility function $U(., \tau_i)$, which is driven by our set point SINR target.

In the following, u_i is the transmit power of the ith network node, and $r_{i,k}$ and $s_{i,k}$ are the RSSI and SINR reported by the ith backlogged mobile. We have chosen these two metrices because they are directly available from most existing WiFi chipsets; in some cases SINR is replaced by the channel quality, which can, however be converted back to SINR units. The noise as measured by the node is given by $N_i = \frac{\sum_k s_{i,k} - r_{i,k}}{\sum_k I_{s_{i,k} > \tau_i}}$ The state transition diagram is given in I. The values ∇, T_s are user provided and can be adjusted depending on the type of the network. As can be seen here, the states *Stable*, *Ramping* and *Backoff* are rough analogues of the different phases of the TCP connection. A separate procedure is implemented for measuring the 'background' noise threshold; in our case, we simply measured the average signal energy of the system when the AP wasn't transmitting. However, specific air interfaces may support direct noise measurement; in defined guard periods, for example, which provide a good approximation of the current system noise.

V. Design and verification

A. Simulation results

Initially, the algorithm was verified in a customized simulation environment. The simulation setup allows us to test the algorithm with a very large number of network node and UE combinations. N network nodes cater to U user devices placed randomly in a fixed area; each network node has a dynamic transmission range of 1dBW to 20dbW and a startup transmit power of 10dbW. User devices attach to the network nodes using measured RSSI and then receive downlink data. The data transmission rates are derived from the measured SINR, so as to cause a BER of 0.01%. Each user equipment receives a random amount of data drawn from a Pareto distribution. Once the data queue goes idle for a user, it can select a new device to campon using either RSSI measurements or a combination of load and SiNR measurements. The setpoint τ_i is configured externally.

The graph in Fig. 2 below shows a comparision of the same system for different combinations of N and U; in one case active interference management is being carried out using the algorithm described above and in the other, there is no interference management. Each line shows the percentage improvement in average throughput for a given number of network nodes (depicted by the variable *Nn*) for the SON algorithm, versus the baseline average throughput when no algorithm is used. The X-axis shows the number of user terminals (distributed randomly over a unit square space) for a given simulation with a particular value of *Nn*. The Y-axis shows the percentage difference in throughput. We can see that the interference management algorithm easily outperforms the baseline when the system is relatively uncrowded and gradually degenerates to the baseline performance as the number of user devices per network node increases.

B. Real life results

We subsequently have implemented and tested our algorithm on a laboratory setup comprising of 8 access points and 16 Wifi clients, all operating on a single WiFi channel in a closed and sanitized laboratory environment. The WiFi network nodes were Linux workstations with attached WiFi cards; the SON algorithm was implemented in an application which controlled the WiFi driver (for power measurements and power settings) using the standard Linux interface. The user terminals were a mixture of commercially available laptops and WiFi enabled mobile phones; the test was carried out in a controlled lab atmosphere and loading was generated using a mixture of artificially generated traffic and smartphone applications, such as web-browsing and game apps. The overall network was monitored using a mixture of inhouse tools and a commercial tool called Ekahau, which dynamically measures network RSSI and SINR from multiple positions.

The results are captured in the graphs in Fig. 3a and Fig. 3b. Since the network setup was fairly dense and the APs were power capped, the average throughput for APs in the

 ISBN: 978-1-61208-279-0

TABLE I. Congestion management - state/event matrix

Current State	Measured event	Action	New State
Stable	Noise measurement is stable or reducing for T_s time periods	tx pwr increases by ∇	Ramping
Stable	Noise measurement is crosses threshold	Reduce tx power to last known stable value	Backoff
Stable	Noise baseline has changed	Adjust tx power accordingly	Stable
Backoff	Noise measurement is stable or reducing for T_s time periods	Save current tx power as last stable tx power	Stable
Ramping	Noise measurement changes by less than ∇	Adjust tx power	Stable
Ramping	Noise threshold changes	Adjust tx power	Stable
Ramping	Noise measurement crosses threshold	Reduce tx power to last known stable tx pwr	Backoff

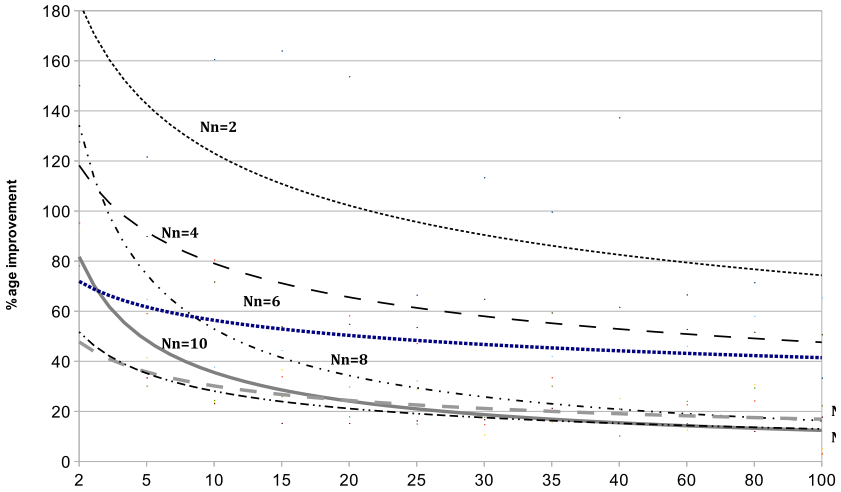
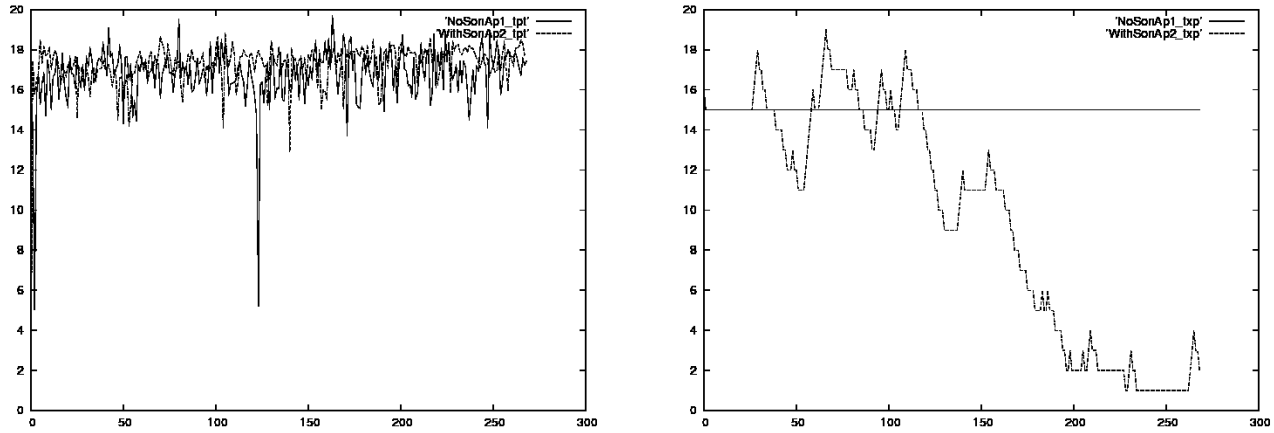


Figure 2. Simulation output; throughput improvement using distributed SON algorithm



(a) AP 1 (b) AP 2

Figure 3. Real life - setup and results

SON and no SON environment were very similar. However, the APs in the network with SON active ran at *substantially* lower transmit power, a full 3-6 dB below the ones with no SON; even though both sets of APs started at the exact same transmit power settings. This arises from the conservative behaviour of the SON algorithm (borrowed from the original TCP). As the number of APs are reduced, the contention drops and individual APs start scaling their power adaptively, leading to substantial improvements in average throughput.

A further observation was that the set point τ_i plays a very important role in the stability of the algorithm. A high value of τ_i actually acts as a damping factor, because all network nodes tend to converge slowly towards this. On the other hand, a low set point allows network nodes to be more responsive to load conditions (since all network nodes achieve the set point easily), at the cost of network fairness;

 ISBN: 978-1-61208-279-0

there is substantial variations in the mean throughput achieved by different network nodes. A second metric is the relative percentage of time a network node spends in stable state; it can be seen that there are substantial variations in this when τ_i is reduced and it reduces as the value of τ_i increases.

VI. Conclusion and Future Work

In this paper, we have formulated a simple, yet robust and generic distributed algorithm for network wide optimization of transmit power levels for a self-optimizing network of WLAN APs. The principles can, in general, be translated to other networks such as LTE - in fact, any network which is power limited can use the algorithm. Simulation and real life results show that our algorithm provides significant improvements over the unmanaged network.

Our laboratory experiments show that the algorithm is adaptive to network conditions, but can lead to unstable equilibria at times. It also suffers from the reverse of the Stackelberg phenomenon; late starting access points tend to drive the equilibrium of the algorithm away from an optimal equilibrium. Future work focuses on mechanisms to limit this, perhaps using supervisory mechanisms.

References

[1] "Aruba adaptive network management," http://www.arubanetworks.com/products/arubaos/adaptive-radio-management, [Accessed January 12, 2013].

[2] "Smartmesh networking," http://www.ruckuswireless.com/technology/smartmesh, [Accessed January 12, 2013].

[3] W. G. 3, ""3rd generation partnership project: Technical report: Self-configuring and self-optimizing network: Use cases and solutions"," Tech. Rep., March 2003.

[4] N. Ahmed and S. Keshav, "Smarta: a self-managing architecture for thin access points," in *Proceedings of the 2006 ACM CoNEXT conference*, ser. CoNEXT '06. New York, NY, USA: ACM, 2006, pp. 9:1–9:12. [Online]. Available: http://doi.acm.org/10.1145/1368436.1368449

[5] I. Viering, M. Dottling, and A. Lobinger, "A mathematical perspective of self-optimizing wireless networks," in *Communications, 2009. ICC'09. IEEE International Conference On*, June 2009, pp. 1–6.

[6] B. Kauffmann, F. Baccelli, A. Chaintreau, K. Papagiannaki, and C. Diot, "Self Organization of Interfering 802.11 Wireless Access Networks," INRIA, Rapport de recherche RR-5649, 2005. [Online]. Available: http://hal.inria.fr/inria-00070360

[7] S. Bhaumik, G. Narlikar, S. Chattopadhyay, and S. Kanugovi, "Breathe to stay cool: adjusting cell sizes to reduce energy consumption," in *Proceedings of the first ACM SIGCOMM workshop on Green networking*, ser. Green Networking '10. New York, NY, USA: ACM, 2010, pp. 41–46. [Online]. Available: http://doi.acm.org/10.1145/1851290.1851300

[8] T. Moscibroda, R. Chandra, Y. Wu, S. Sengupta, P. Bahl, and Y. Yuan, "Load-aware spectrum distribution in wireless lans," in *Network Protocols, 2008. ICNP 2008. IEEE International Conference on*, vol. 3, no. 3, October 2008, pp. 137 –146.

[9] D. Divsalar, "A simple tight bound on error probability of block codes with application to turbo codes," *Journal of Programming Languages*, November 1999.

[10] P. Gupta and P. Kumar, "The capacity of wireless networks," *Information Theory, IEEE Transactions on*, vol. 46, no. 2, pp. 388 – 404, March 2000.

[11] D. Qiao, S. Choi, and K. G. Shin, "Interference analysis and transmit power control in ieee 802.11a/h wireless lans," *IEEE/ACM Transactions on Networking*, vol. 15, no. 5, pp. 1007–1020, October 2007.

[12] K. Kowalik, M. Bykowski, K. B, and D. M, "Practical issues of power control in ieee 802.11 wireless devices," in *IEEE International Conference on Telecommunications (ICT 2008), Proceedings of*, June 2008.

[13] L. Kleinrock and B. Tung, "Distributed control methods," in *Proceedings of the 2nd International Conference on High Performance Distributed Computing*, July 1993, pp. 206 – 215.

[14] M. Mathis, J. Semke, J. Mahdavi, and T. Ott, "The macroscopic behaviour of the tcp congestion avoidance algorithm," *SIGCOMM Computer Communication Review*, vol. 27, no. 3, July 1997.

[15] L. Brakmo and L. Peterson, "Tcp vegas: end to end congestion avoidance on a global internet," *Selected Areas in Communication, IEEE Journal on*, vol. 13, no. 8, pp. 1465–1480, October 1995.

[16] H. Balakrishnan, V. PAdmanabhan, S. Seshan, and R. Katz, "A comparision of mechanisms for improving tcp performance over wireless links," *Networking, IEEE/ACM Transactions on*, vol. 5, no. 6, pp. 756–769, December 1997.

[17] J. Mo, R. La, V. Ananthanram, and J. Walrand, "Analysis and comparision of tcp reno and vegas," in *IEEE Infocom, 99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3, March 1999, pp. 1556–1563.

[18] S. H. Low, "A duality model of tcp and queue management algorithms," *IEEE/ACM Transactions on Networking*, vol. 11, no. 4, August 2003.

[19] S. Kunniyur and R. Srikant, "End to end congestion control schemes: Utility functions, random drops and ecn marks," *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, October 2003.

 ISBN: 978-1-61208-279-0

SLA Framework Development for Content Aware Networks Resource Provisioning

Cristian Cernat, Eugen Borcoci, Vlad Poenaru
Telecommunication Department
University POLITEHNICA of Bucharest
Bucharest, Romania
cristian.cernat@elcom.pub.ro, eugen.borcoci@elcom.pub.ro, vlad.poenaru@elcom.pub.ro

***Abstract* — This paper develops a Service Level Agreement (SLA) framework as a part of a complex networked media multi-domain eco-system, aiming to deliver multimedia Quality of Services (QoS) enabled services over multiple domain network infrastructures. It continues a previous work which defined the management architecture of a networked media oriented system - based on new Virtual Content Aware Networks (VCAN) concepts. The contribution of this work, beyond traditional SLA usage, consists in fully specifying and then implementing the dynamic SLA-based management for VCAN resources provisioning in a multi-provider, multi-domain network environment.**

Keywords — Content-Aware Networking; Service Level Agreement, Multi-domain; Management; Resource provisioning; Future Internet.

I. INTRODUCTION

A strong orientation towards content/information is expected for the current and Future Internet (FI), estimating that content distribution will cover approximately 90% of the total traffic in 2015 [1-4]. Among new architectural concepts proposed, there are Content-Awareness at Network layer (CAN) and Network-Awareness at Applications layers (NAA). This approach can create a *cross-layer optimization loop* between the transport, applications and services, which did not exist in traditional layered architectures and also better serves the content/information centric trends, [4].

The European FP7 ICT research project, "Media Ecosystem Deployment Through Ubiquitous Content-Aware Network Environments", ALICANTE, [5], adopted the NAA/CAN approach. It defined, designed and currently is implementing this complex multi-domain Media Ecosystem. The work described here is a continuation of the research work associated with a part of this project. That is why, the complete system description and main system level architectural and design decisions cannot be detailed here; references are indicated to help the reader to get a more complete view on the system.

Several cooperating environments and business actors are defined in [5]: *User Environment* (UE), containing the End-Users; *Service Environment* (SE), containing Service Providers (SP) and Content Providers (CP); *Network Environment* (NE), comprising a new business entity called CAN Provider (CANP) and the traditional Network Providers (NP). An NP manages the network elements, in the traditional way, at IP level. The "environment" is a generic grouping of functions working for a common goal and which possibly vertically span one or more several architectural (sub-) layers.

This paper further specifies and develops a SLA – based framework to create and dynamically provision VCANs. The VCANs in ALICANTE approach can be seen as a light virtualization solution, i.e. parallel multi-domain logical Data planes, each one dedicated to certain classes of services and types of media flows. This approach is similar to that presented in [9] but enriched here with content awareness and QoS capabilities. The CAN Provider creates VCANs at request of the Service Provider which negotiate SLAs with CANP. Inside each VCANs the QoS assurance is performed by several mechanisms: static and dynamic provisioning, transport service differentiation and media flow dynamic adaptation.

The paper continues the work presented in other related papers such as [13-15] with focus on SLA framework as a part of the service and resource management. In addition to the previous work, *this paper develops the complete design, implementation and validation of the solution.* The paper structure outline follows. Section 2 presents related work. Section 3 provides a short overview of the system architecture. Section 4 develops the full SLA framework. Section 5 presents samples of implementation solutions. Conclusions and perspectives are outlined in Section 6.

II. RELATED WORK

Complex communication systems involve several business entities playing providers and customers roles. Usually the relationships between such entities are described by SLAs, [5-10].The technical part of the SLA is frequently named *Service Level Specification (SLS)*. While SLA have been proposed in many studies, their usage especially in a dynamic context is still an open research topic.

In MESCAL project, [6] the SLS concepts are extended to cover multiple-domain networks. It is defined a *Service Provider* actor dealing with higher level services and *IP Network Providers* offering connectivity services. Additionally, the concept of Meta QoS class is introduced in

 ISBN: 978-1-61208-279-0

[7], as a set of well known identifiers and parameter ranges for a few number of service classes, covering all significant application flows types, aiming to simplify the inter-domain signaling related to QoS composition. In [8] a system for media distribution over multiple heterogeneous network domains is proposed. The business model includes several actors: *Service Provider (SP), Network Provider (NP), Content Consumer (CC), Content Provider (CP), Access Network Provider (ANP).* Several types of SLA/SLSs exist customized to fulfill the needs of media related applications and services. However, all three solutions presented above consider only a single logical network infrastructure serving different traffic flows.

A new concept of *Parallel Internets (PI)* is proposed in [9], to enable end-to-end (E2E) service differentiation across multiple administrative domains, based on logical slicing of the Internet. The PIs can coexist, as logical networks composed of interconnected, per-domain, Network Planes (NPl). The actor types are Service Provider (SP), IP network provider (INP) and customers. INPs offer IP connectivity services and SPs offer high level services to customers. Several types of SLAs are defined, and also INP interconnection agreements (NIAs). However, no content awareness concept is present in [9].

The FP7 project COMET *"COntent Mediator architecture for content-aware nETworks",* [10] provides a *unified* interface for content access whatever the content characteristics are: *temporal nature*, *physical location*, *interactivity requirements*, etc. It enables the most appropriate E2E transport by mapping the content (according to its requirements and user preferences) to the appropriate network resources. The objective is to get the best quality of experience (QoE) for end users; it supports unicast, anycast and multicast. A Content Mediation plane is introduced between ISPs and content servers, combining content resolution and access: locating content according to delivery requirements (content mediation); delivering it using the most suitable resources (network mediation). SLAs are defined between the business entities.

In ALICANTE, [5], [12], a similar concept to parallel virtual planes [9] is adopted, but modified and enriched with *content awareness*. There will be (generally) a one-to-one mapping between a network data plane and a VCAN - where each VCAN is customized for a given class of services and type of media flows. The COMET and ALICANTE have overlapping scopes. However, the COMET business model is only partially sufficient for ALICANTE needs: no powerful user and service environment exists; service environment functions are embedded in ISPs; no CAN Provider exists. COMET does not fully consider the cooperation between network overlay and network resources, but is focused mainly on mediation activities. There is no complete chain of services management. ALICANTE considers additionally the concept of home network and its associated Home Box and exploits CAN concepts and optimization loop between applications/services and network.

In a recent work [11], a content centric solution is proposed. The CURLING, "*Content-Ubiquitous Resolution and Delivery Infrastructure for Next Generation Services*", aims to media content distribution at massive scale. It has a holistic approach, (content publication, resolution and, delivery) and provides to Content Providers and customers high flexibility, to express their location preferences when publishing and requesting content, respectively - through *scoping* and *filtering* functions. Business relationships are defined between ISPs, including local ISP policies, and specific CP and customer preferences. ALICANTE is complementary to CURLING; it is less content centric in the control plane, but more powerful in assuring efficient media flow QoS- enabled transport, based on content awareness. The SLA framework is very flexible allowing dynamic creation, modifications and termination of VCANs.

III. CONTENT AWARE NETWORKS –SYSTEM ARCHITECTURE

The ALICANTE concepts and architecture have been defined in [5], [12], where main selection of the architectural solutions and design decisions are motivated. The connectivity services management architecture is described in [13]. Figure 1 shows a simplified picture, of the management and control (M&C) plane. The network contains several Core Network Domains (CND), belonging to NPs, and also access networks (AN). The ANs are out of scope of VCANs, given the large variety of technologies and degree of resource management. The CAN layer M&C is partially distributed: one *CAN Manager* (CAN_Mgr) belonging to CANP exists for each IP domain, doing VCAN planning, provisioning, advertisement, offering, negotiation, request for installation and exploitation. Each network domain has an *Intra-domain Network Resource Manager* (Intra_NRM), as the ultimate authority configuring the network nodes. This architectural solution allows an incremental deployment for Network Providers: an NP can be enhanced in order to become also a CAN Provider. The End User (EU) terminals are connected to the network through Home Boxes (HB). The novel CAN routers (not shown here) are called *Media-Aware Network Elements (MANE)* to emphasize their main additional capabilities: content and context – awareness. The CAN layer cooperates with HB and Service Environment by offering them CAN services.

The architectural solution for VCAN Management has been already defined in [13]. Here only a short summary is recalled for sake of clarity. A functional block at SP performs all actions needed for VCAN support (planning, negotiation with CANP, VCAN exploitation) while the CAN Manager at CANP performs VCAN provisioning and operation. The two entities interact based on the SLA/SLS contract initiated by the SP (SLA1 – in the Figure 1). Then the initiator CAN manager (e.g., CAN Manager 1) has to discuss with other involved CAN Managers in hub style (the topology discovery and determination of the other domains involved in a multi-domain VCAN are not subjects of this paper) negotiating with them SLAs (e.g., SLA2.1, SLA 2.2).

 ISBN: 978-1-61208-279-0

Each CAN Manager at its turn has to ask resources from its associated Intra_NRM and establishes with it a SLAs (SLA3.1, SLA3.2 and SLA3.3). To do this, CAN Manager runs a *combined routing, reservation and VCAN mapping algorithm* described in [15]. If the set of SLAs is successful, then the Intra_NRMs receive commands from their CAN Managers to install the configurations in routers, as to assure the QoS characteristics of the VCANs. These actions can be done immediately after SLA agreements or later (conforming to the options expressed in SLA1). Later, the media flow will be transported through the QoS enabled VCAN pipes from content servers to end users. The network technologies supporting these are Multiprotocol Label Switching (MPLS) and/or differentiated services (DiffServ, [5]).

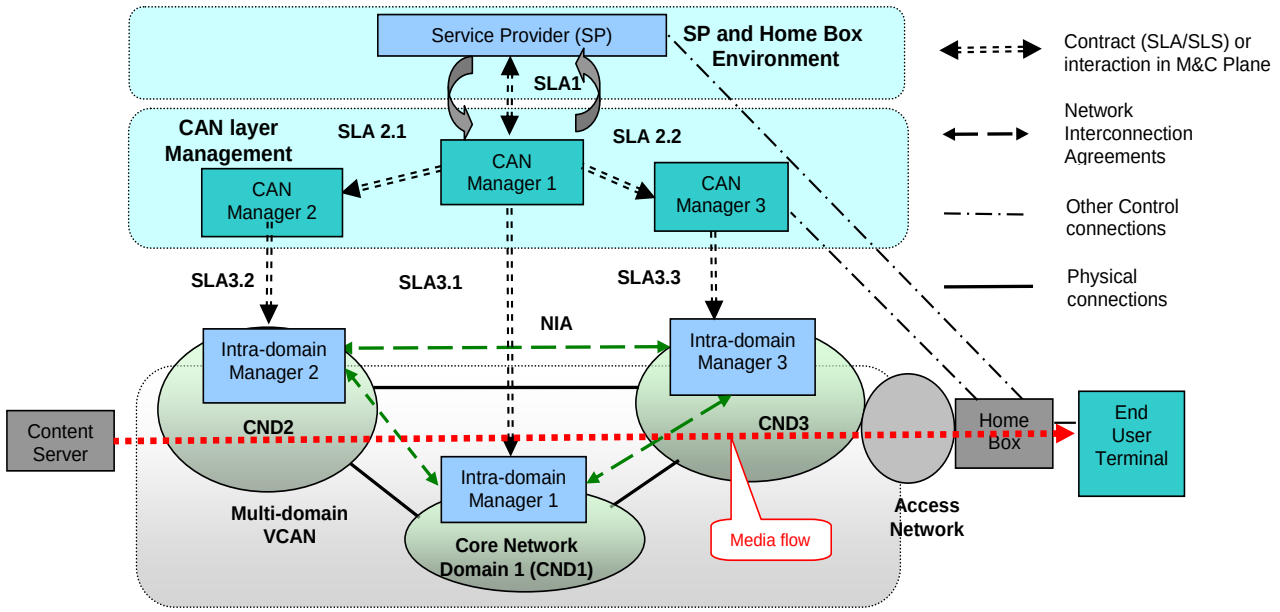


Figure 1. Business Actors and SLA framework high level view

IV. SLA/SLS FRAMEWORK FOR RESOURCE PROVISIONING

The general ALICANTE template of an SLS between the SP and CANP has been defined, able to support more or less sophisticated SP requirements (in practice only a sub-set of the parameters might be specified by the Service Provider).

Table 1 shortly describes the sections of an SLS between an SP and the initiator CAN Manager. Unspecified SLS values (by the SP) let CAN Manager and/or Intra_NRM to apply their own policies. The SLS is transported through negotiation protocols (vertically between SP-CAN Manager and horizontally between initiator CAN Manger and other CAN Managers of the involved core network domains.

TABLE I. TYPICAL SLS (SP-CANP) TEMPLATE EXAMPLE

SLS sections	SLS Element/Clause
General	1. SLS Identification 2. VCAN associated CATI (Content Aware Transport Information) optionally inserted in the data packets by the content servers describes Service Type, Service sub-type, etc.
VCAN Connectivity services requirements needed by SP for VCAN	1. Topology (pipe, hose, funnel) and scope (ingress, egress points). Identifies topology and the edge points of the topological region over which the QoS applies. 2. Connectivity class quantitative/qualitative: Bandwidth, Delay, Jitter, Loss, Availability. 3. VCAN time life, optionally define the time-life of this SLS.
VCAN Traffic Processing Requirements	1. Traffic identification: Ingress flow Id, Egress flow Id, Ingress point, Egress point. 2. QoS guarantees conformance algorithm: Traffic Control (TC) algorithm and parameters. 3. DiffServ-like treatment of excess traffic: dropping, re-marking, shaping, adapting. 4. Routing and forwarding rules: constraints on the way to compute the paths for forwarding. 5. Adaptation requirements: describe the condition (thresholds, etc.), under which the media flow adaptation is allowed for Partially Managed and Unmanaged ALICANTE services.
VCAN Services	1. Monitoring methods: specifies how the SLS should be verified at CAN layer.

 ISBN: 978-1-61208-279-0

Assessment Requirements	2. Monitoring tasks: describes when and how the monitoring actions are performed. 3. Details of notification and reports: time, level of information aggregation, etc.
VCAN Allowed actions	1. Availability and VCAN service schedule: time intervals allowed for service invocation. 2. Invocation methods: conditions of invocation (see also the time life of a VCAN). 3. Modification permission of Connectivity services, Traffic Processing and Service Assessment: describe the modifications allowed for the three categories.

V. DEVELOPMENT AND IMPLEMENTATION RESULTS

This section will present the main implementation decisions for SLA framework and samples of results.

A. SLS signaling

Several solutions and scenarios have been analyzed for different cases of VCAN realization. Two step approach has been selected for VCAN construction: *negotiation and implicit resource reservation; VCAN installation in the network routers.* Usually the Service Provider initiates the new VCAN construction. After negotiation and logical resources reservation the SP requests installation (invokes) this VCAN, i.e. asks to CAN Manager (and this asks to Intra_NRM) to install the VCAN appropriate configurations in the routers. A single SLS per VCAN, or several ones, can be negotiated if these SLSs define the same QoS and security class.

One basic signaling case (described below) is for a unicast communication VCAN, where the virtual VCAN pipes are implemented on top of MPLS+Diffserv enabled paths. To decide the final implementation solution several signaling variants have been analyzed and compared.

a. Information on network resources (controlled by Intra-NRM) are obtained by the CAN Manager on demand, i.e. asynchronously, when a new SP request arrives. The advantages are: network availability updates in the CAN_Mgr Data Base are done only when needed, i.e. at each new SP request (for initiator CAN_Mgr), or at request of other CAN_Mgrs. The Intra_NRM does not care to inform CAN_Mgr each time when it makes its *Resource Availability Matrix (RAM)* modifications. However, a higher signaling overhead exists: each time that a new request comes to this CAN_Mgr, from SP or other CAN_Mgrs, the RAM information should be asked for each domain.

b. *Network RAM delivered proactively to CAN Manager, asynchronously, by Intra_NRM*.

The RAM is proactively uploaded to CAN_Mgr at Intra_NRM initiative, (when modifications appear, or periodically).. Less signaling is needed when new SP requests (or request from other CAN_Mgrs) arrive at this CAN_Mgr (the RAM information is already available). However, tighter synchronization Intra_NRM – CAN_Mgr, is necessary, given that Intra_NRM might change RAM info while CAN_Mgr is solving some computation based on previous RAM values.

In the current implementation, the solution *on demand* has been selected. Figure 2 presents the signaling actions implemented in case of a multi-domain VCAN requested by the SP to CAN_Mgr1 (VCAN initiator). It has been supposed that CAN_Mgr2 and 3 should participate to the VCAN construction, spanning three core network domains. Similar diagrams have been designed and verified in implementation, e.g. for multicast VCANs, [15].

c. An advanced solution is necessary when different VCANs have different life-time; the resource reservation should consider such intervals. To simplify the management, a time-unit T can be defined and a VCAN time-life can be kT, with k natural number. The VCAN construction and/or reconfiguration happens at nT instants only (the T value can be selected by the Can Manager policy - e.g., few minutes). At each nT instant one can update the RAM information in CAN Manager, or asynchronously (at Intra_NRM initiative) when major changes happens to its RAM.

The implementation technology used in signaling framework is shortly described below. The CAN Manager has been implemented by using *Python* language. This allows to deploy the applications on multiple Operating Systems and to have a self contained system that can be installed without affecting the rest of the applications installed. *Virtual environments* technology that comes with Python has been used. For database access we use an ORM (Object Relational Mapper), *SqlAlchemy*, which helps to hide the details while having the full power to interact with the data.

The CAN Manager interaction to other entities (SP, Intra_NRM, or other CAN Managers) is done using web services. Several frameworks for SOAP have been analyzed; *Suds* has been selected for initiating web service requests and *Spyne* for listening to requests. The application is presented as a *Web Server Gateway Interface* (WSGI) object, a Python protocol standard for web applications, and can be deployed on any WSGI capable container. We are mainly using *Spawning* as a container started as a service at system startup, but we also tested on the *nginx web server* with *uWSGI* as container. The OS we deployed this on is a Debian distribution of GNU/Linux.

The CAN Manager listens for SP requests and then continues by running its internal algorithms. After computing a VCAN mapping for the input RAM, it sends the requests to all other CAN Managers involved in the VCAN. In the sense of SOAP specification the CAN Manager has two listening ports: one for SP requests and another one for other CAN Managers requests. The CAN Manager may issue requests to either of two entities: the Intra_NRM, and other CAN Managers. The database is accessed using database specific protocols.

A possible implementation improvement is to make asynchronous all calls between entities. However this involves a lot of glue multi-threading code; so for the current implementation synchronous calls have been selected.

 ISBN: 978-1-61208-279-0

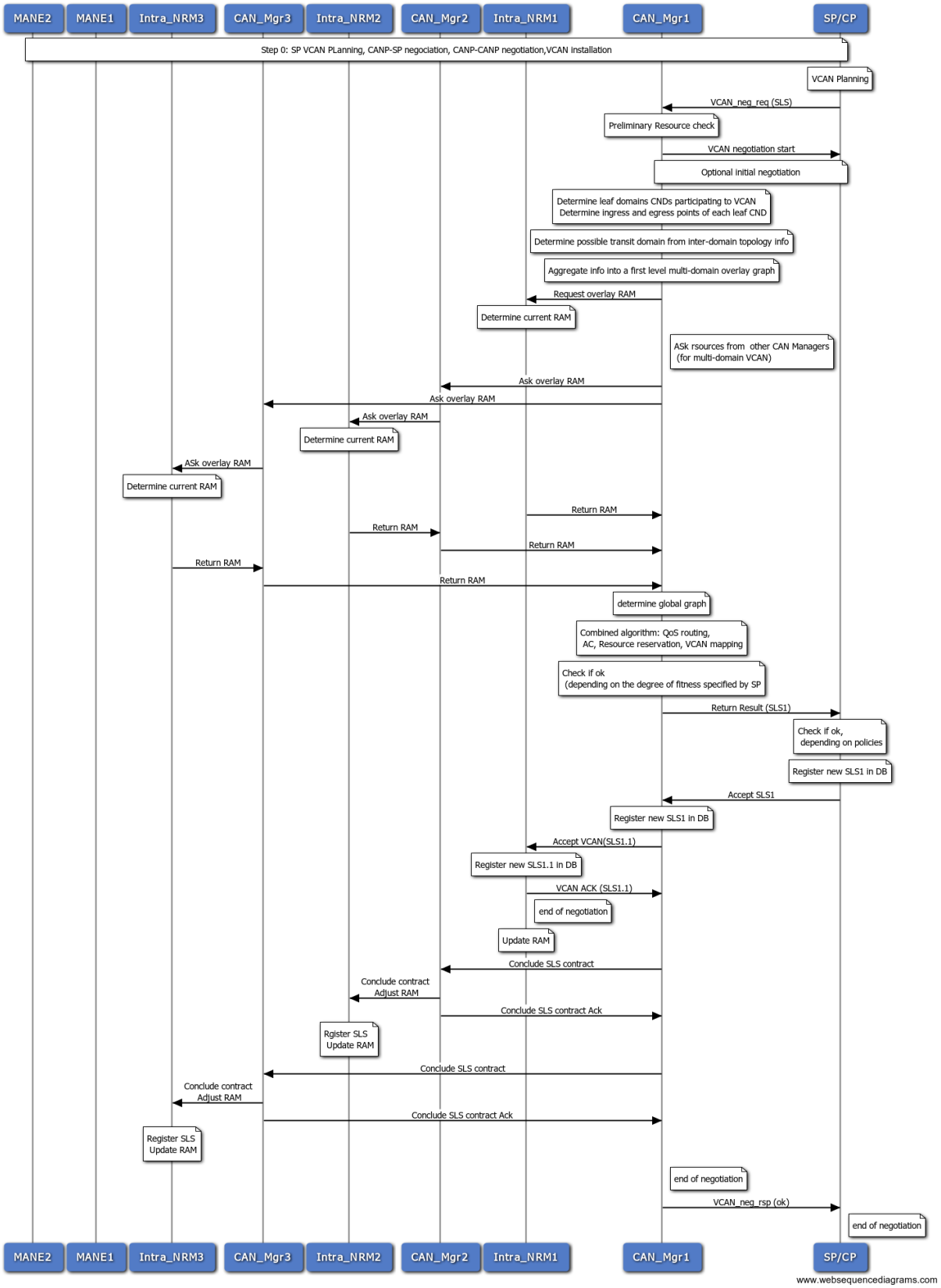


Figure 2. Implementation solution of the signaling for multi-domain unicast VCAN (MPLS based)

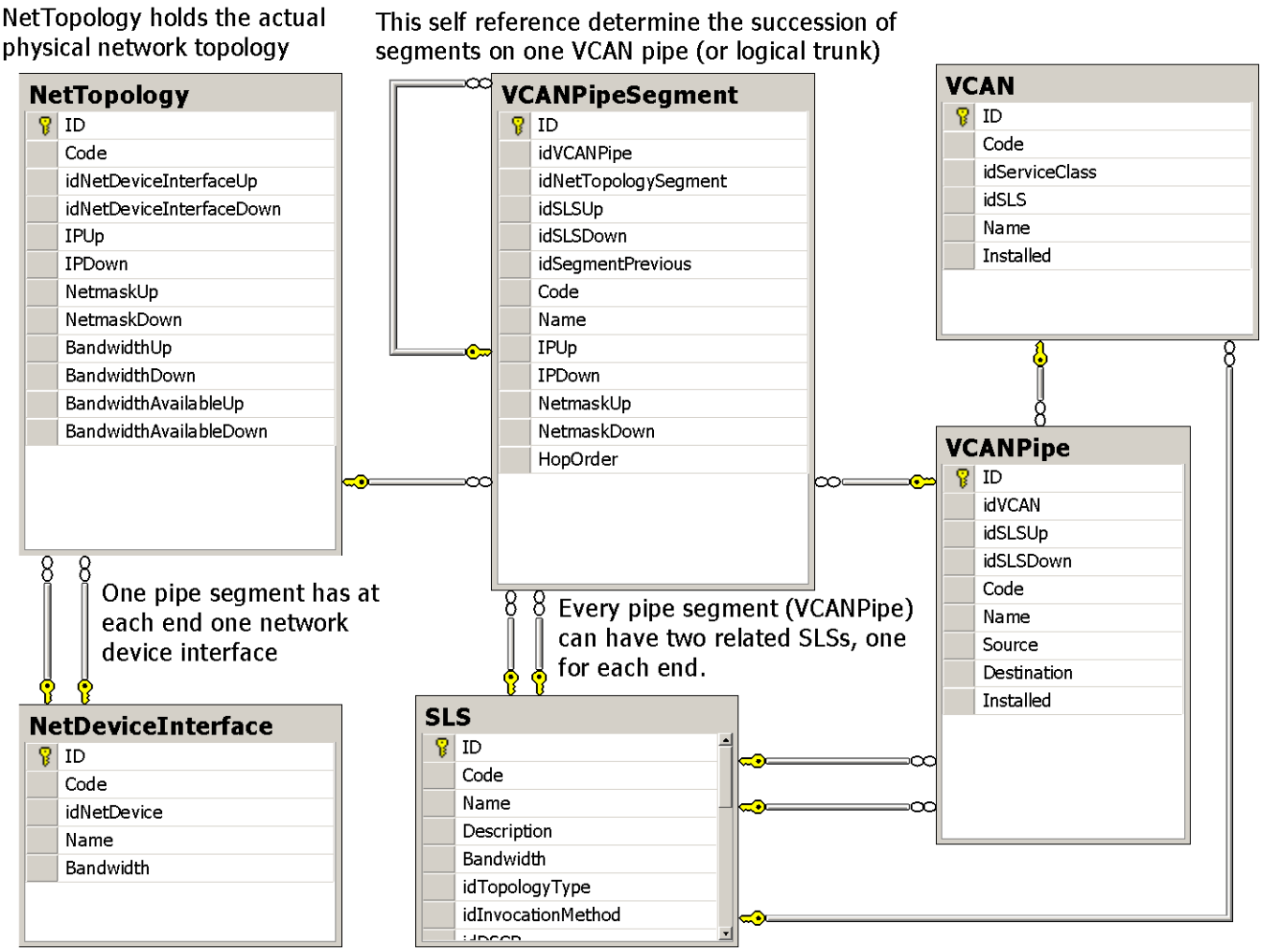


Figure 3. Sample of the Data Base at CAN Manager: Table *VCANPipe* and its connections with neighbor tables

B. *Data Base at CAN Manager*

A relational database has been designed and implemented around several main building blocks or tables: *"SLS"* for SLS templates used to hold negotiated resources around the network, *"NetTopology"* holds the image of actual physical network topology on which VCAN pipelines are built, *"VCANPipeSegment"* for VCAN Pipe Segments and *"NetDeviceInterface"* for information about network path endpoints known by the entity using this database (SP or CAN_Mgr). It is not mandatory for an endpoint to represent the actual physical interface located on network devices.

Figure 3 presents an overview of the current Database implementation that holds only basic SLS parameters (like *Bandwidth*, *Delay*, *Loss* or *Jitter*) in the main SLS table, other parameters suitable in some category will be hosted by satellite tables having prefix "SLS_". There are a couple of reasons for this design: it is much easier to further expand the list of SLS parameters compared to the other design where everything is stored in one big table, better readability of the data structures for someone who knows the design documentation, hierarchical organization.

CAN Manager database stores and provides access to an overlay image of the provisioned MPLS pipelines. This is translated into a matrix of the managed network domains by Intra_NRM having vector elements of type *{..., (input_router_addr, output_router_addr, Bandwidth, x, y), ...}* where parameters x and y are intended for future usage. Every vector element is the concrete representation of one MPLS Label Switched Path (LSP) with a minimum of information as terminal end points and assigned bandwidth.

Data Base contains four different table types according to their role in the architecture; therefore the stored information is specific to some: *Entity (SLS, VCAN, VCANPipe, NetDevice, NetDeviceInterface)*, satellite information related to some entity (*SLS_QoS, SLS_ServiceClass, SLS_CATI*), classification items that enumerate the available entity types (*NetDeviceType, TopologyType*), relationship between entities (*VCANxSLS, SLSxCATI*). The Data Base at CAN_Manager has an interface with the external world (i.e. other functional modules of the CAN_Manager or even

 ISBN: 978-1-61208-279-0

external); for this reason an embedded application programming interface (API) has been implemented that provides CRUD (*create*, *read*, *update and delete*) functionality, data validation, automation procedures, path computations, activity logging and more. In terms of security, the Data Base design provides some advantages in hiding important information about VCAN network to the outside world.

All active modules (including the VCAN – related algorithms, i.e. routing, resource reservation and mapping) and Data Base have been implemented. The message sequence chart presented in Figure 2 and other similar scenarios (e.g. for multicast VCANs) have been validated.

VI. CONCLUSIONS AND PERSPECTIVES

This paper presented an SLA/SLS framework development for dynamic Virtual Content Aware Network (VCAN) provisioning, spanning multiple core network domains. The SLAs are customized as to satisfy Service Providers requirements for media distribution services.

The architectural and implementation solutions presented have several advantages. A light virtualization solution for multi-domain media distribution QoS enabled is offered, creating a cross-layer optimization loop between the transport and application layers. The Intra-domain network resource manager independence (important business requirement) is preserved, due to two-level hierarchical architecture (CAN and network layers). All tasks to construct multiple domain VCANs are delegated to CAN Provider (Service Providers are just asking for VCANs and then using them). Seamless deployment is possible: VCAN guaranteed services can coexist with best effort ones (families of SLS templates can be defined). The CAN Manager software can be naturally installed as an upgrade of the Intra_NRM, thus transforming an NP in a CAN Provider.

The management system is scalable given that the VCANs are installed as "networks", on demand, on short-mid-long term, and not at per-flow demand; therefore the signaling tasks do not create a significant overhead.

The architecture is similar to the new approach of Software Defined Networking (SDN) [16], [17] where important control functions are moved out of data plane into a separate control plane. This decoupling enables both planes to evolve independently, and brings advantages such as high flexibility, vendor-agnosticism, programmability, and the possibility of realizing a centralized network view.

The implementation of the proposed system is in final phase in the ALICANTE project. Future work may envisage the study of a migration towards a solution closer to SDN.

Acknowledgments

This work was supported partially by the EC in the context of the ALICANTE project (FP7-ICT-248652) and partially by the projects POSDRU/107/1.5/S/76909.

REFERENCES

[1] J. Schönwälder, M. Fouquet, G .Dreo Rodosek, and I.C Hochstatter, "Future Internet = Content + Services + Management", IEEE Communications Magazine, vol. 47, no. 7, Jul. 2009, pp. 27-33.

[2] T. Anderson, L. Peterson, S. Shenker, and J. Turner, "Overcoming the Internet Impasse through Virtualization", Computer, vol. 38, no. 4, Apr. 2005, pp. 34–41.

[3] 4WARD, "A clean-slate approach for Future Internet", http://www.4ward-project.eu/ (last access May 2012).

[4] 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.

[5] ALICANTE, Deliverable D2.1, ALICANTE Overall System and Components Definition and Specifications, http://www.ict-alicante.eu, Sept. 2011.

[6] M. P. Howarth, P. Flegkas, G.Pavlou, N.Wang, P.Trimintzios, D.Griffin, J.Griem, M.Boucadair P.Morand,A.Asgari, P.Georgatsos, "Provisioning for Interdomain Quality of Service: the MESCAL Approach", IEEE Communications Magazine, June 2005, pp.129-137.

[7] P. Levis, M. Boucadair, P. Morrand, and P. Trimitzios, "The Meta-QoS-Class Concept: a Step Towards Global QoS Interdomain Services", Proc. of IEEE SoftCOM, Oct. 2004.

[8] ENTHRONE, End-to-End QoS through Integrated Management of Content, Networks and Terminals, FP6 project, www.ist-enthrone.org/ (last access May 2012).

[9] M. Boucadair, P. Lévis,D. Griffin, N. Wang, M. Howarth, G. Pavlou, E. Mykoniati,P. Georgatsos, B. Quoitin, J. Rodríguez Sánchez, M. L. García-Osma "A Framework for End-to-End Service Differentiation: Network Planes and Parallel Internets", IEEE Communications Magazine, Sept. 2007, pp. 134-143.

[10] F.J. Ramón Salguero, "COntent Mediator architecture for content-aware nETworks", 2010, http://www.comet-project.org/.

[11] W.K. Chai, et.al., CURLING: Content-Ubiquitous Resolution and Delivery Infrastructure for Next-Generation Services , IEEE Communications Magazine, March 2011.

[12] H. Koumaras , D. Negru , E. Borcoci , V. Koumaras , C. Troulos , Yes. Lapid , E. Pallis , M. Sidibé ,G. Gardikis, G. Xilouris, C. Timmerer, „Media Ecosystems: A Novel Approach for Content-Awareness in Future Networks" FIA Book 2010

[13] E. Borcoci, M. Stanciu, D. Niculescu, D.Negru, G. Xilouris, "Connectivity Services Management in Multi-domain Content-Aware Networks for Multimedia Applications " , Proc. of. INTERNET 2011, Luxembourg, June 2011, www.iaria.org.

[14] E.Borcoci, G.Carneiro, R. Iorga**,** "Hybrid Multicast Management in a Content Aware Multidomain Network**",** AFIN 2011, The Third International Conference on Advances in Future Internet, http://www.thinkmind.org/index.php?view=article&articleid=afin_2011_5_20_70107

[15] E.Borcoci, R.Miruta, S.Obreja, "Multi-domain Virtual Content-Aware Networks Mapping on Network Resources", EUSIPCO Conference, 27-31 August 2012, Bucharest, http://www.eusipco2012.org/home.php

[16] S.H.Yeganeh, A.Tootoonchian, Y.Ganjali, "On Scalability of Software-Defined Networking", IEEE Communications Magazine, February 2013, pp.136-141.

[17] B.Raghavan, T.Koponen, A.Ghodsi, M.Casado, S. Ratnasamy, S.Shenker, "Software-Defined Internet Architecture: Decoupling Architecture from Infrastructure", HotNets-XI Proceedings of the 11th ACM Workshop on Hot Topics in Networks, 2012, pp. 43-48, doi: 10.1145/2390231.2390239.

 ISBN: 978-1-61208-279-0

Modelling Mobility-Aware Applications for Internet-based Systems

Bruno Yuji Lino Kimura
Instituto de Matemática e Computação - IMC
Universidade Federal de Itajubá - UNIFEI
Itajubá - MG, Brasil
Email: kimura@unifei.edu.br

Edson dos Santos Moreira
Instituto de Ciências Matemáticas e de Computação - ICMC
Universidade São Paulo - USP
São Carlos - SP, Brasil
Email: edson@icmc.usp.br

***Abstract*—Future network architectures, such as Next Generation Network (NGN) and Vehicular Communication Network (VCN), will provide large scale mobile Internet access supported with multiple and small cells of wireless communication. Therefore, the access to the Internet services are prone to very frequent disconnections when the mobile node migrates across the cells. This paper deals with this concern and discusses issues involved on the modelling of mobility-aware applications. An abstract model is described by means of two Finite State Machines (FSM) designed for both mobile nodes, client and server. The FSMs are meant to networked applications to preserve their integrity in the event of frequent connection disruptions in Single Jumps (client side mobility) and Double Jumps (both client and server sides mobility). The FSMs for mobility-aware applications were evaluated using a handover delay mathematical model. With handover latencies from 230 ms in single jumps and from 450 ms in double jumps, simulation results show that the proposed strategies are capable of providing automatic Transmission Control Protocol (TCP) connection re-establishment with smooth impact when increasing the frame error rate and the data lost in handovers.**

Keywords-Mobility-Awareness; Application Layer Mobility; Internet-based systems; IP Mobility Management.

I. INTRODUCTION

The paradigm of Mobile Computing is already part of our modern lifestyle. However, new challenges are arising with future wireless networks. While NGNs consider numerous heterogeneous wireless networks with small cells to increase scalability, mobile nodes are becoming devices of higher mobility. Recently, Internet-based applications for safety and infotainment have been designed to be deployed in vehicles. These applications suffer frequent connection disruptions while the node migrates quickly among the road-side wireless Internet Protocol (IP) networks. These scenarios demand research efforts to enable efficient support for IP mobility in the TCP/IP protocol stack.

Given the common absence of network support for mobility, several solutions have been proposed to work at the various layers of the TCP/IP protocol stack. Mobile IP [1][2] is the general IP mobility solution in the literature. It is designed to work at the Network Layer to provide the special handling required in the addressing and forwarding configurations when a node is moved. Applications running on the mobile node do not need to deal with mobility. Such transparency is an important requirement for IP Mobility Management.

Considering upper layers (Transport and Application) and lower layers (Link and Network) to classify mobility protocols, the choice of the layer where the mobility handling should be implemented implies on where to put the inherent overhead:

i. *For the sake of transparency, mobility can be handled by the lower layers to reduce the impact on the upper layers*. This is achieved with additional infrastructures for managing location of mobile nodes and agents to forward incoming packets to the current location of the nodes [1][2]. However, specific infrastructure increases the cost of deployment and maintenance of the mobility solution. In addition, agents imply modifications in the core of IP networks.

ii. *Handling mobility at upper layers, then the mobile node takes control of its own mobility*. At the application level, host mobility implies the complete re-establishment of broken connections, peer authentication, and recovering data at the Socket buffers lost with disruptions [3][4]. The cost of deployment is reduced, since the handling is conducted end-to-end, therefore, there is no need for agents. On the other hand, it increases the overhead at the upper layer with the burden of the mobility.

In this sense, including mobility support in the TCP/IP stack is a dilemma. However, upper layer based solutions are less costly [5]: the support is provided without requiring a node to be attached to a Home Network, then there is no need for additional infrastructure; solutions are implemented as software components, so that their deployment and maintenance are easier; route optimization is inherent, since there is no entity to intermediate communicating along the end-to-end path; the mean-time-to-recovery (i.e., handover latency) takes no longer than a second [3], therefore, performance is also an advantage. Due to these reasons, several mobility protocols [3][4][6]–[10] are in favour of handling mobility at the upper layers.

There is a promising trend of designing communication

 ISBN: 978-1-61208-279-0

solutions to work at the Application Layer. With recent Web technologies, e.g., HTML5, in support for new paradigms, such as Cloud Computing and Internet of Things, the Web browsers became a generalized interface to access distributed resources on the Internet. In the meantime, efforts on Software-Defined Networking (SDN) have been made toward IP programmable networks by separating and implementing the traffic control as software service.

We advocate this trend and, in this paper, we discuss an abstract model for handling mobility at the Application Layer. Two Finite State Machines (FSM) are exploited to define distributed systems aware of communication events caused by the host mobility on the Internet. We aim to provide a general purpose mobility solution that can be implemented as a software service according to the computer architecture and language of choice of developers. We evaluated the performance of mobility-aware applications according to the handover delay model discussed in [11] and [12]. Simulation results show good performance when degrading parameters of quality of handovers, such as frame error rate and retransmission of data lost in mobility. Handover delays take no longer than 793 ms for the worst case.

This paper is organized as follows: in the next section , we describe the mobility-aware application model; Section III describes the handover delay mathematical model used to evaluate performance of the proposed mobility-aware applications; Section IV discuss results obtained from simulations using the handover delay model; and the last section brings our conclusions.

II. Mobility-Aware Application Model

We assume that the host mobility causes a failure F that breaks the TCP connections of a networked application. Failure F lasts a disconnection time, which is the handover latency $T_{handover}$, until the mobile node acquires Layer 2 and Layer 3 access at the visited network and the mobility protocol resumes ongoing communications. To provide seamless mobility the disconnection latency $T_{handover}$ should be as short as possible, which is a big challenge on Mobile Computing under research efforts in the last decade.

F is denoted by one of the following events:

$$F = \{\texttt{F:TO}, \texttt{F:RST}, \texttt{F:DU}, \texttt{F:BC}\}, \quad (1)$$

where they represent *connection timeout*, *connection reset*, *destination unreachable*, and *broken connection*, respectively. These events lead the system to unexpected states and, without the proper handling, they make the system crash.

We propose a repair plan at the application runtime to handle F events by means of two Finite State Machines, shown in Figure 1, regarding mobile clients connected to mobile servers. They identify unexpected states as erroneous states (grey circles). When reaching erroneous states, the system is able to lead to a safe state again (white circles) with transition operations based on the semantic of the Socket API, so that the communication is resumed consistently without losses. This system property provides mobility awareness to networked applications.

Next, we discuss fundamental aspects for handling mobility with the proposed model.

A. Session establishment

We assume that the mobile nodes are uniquely identified by an identifier decoupled from the IP address, as in [10]. IP addresses become locators that indicate the current logical location of the mobile nodes. When the (re)connection is established, we suggest a handshake for the nodes exchanging their local control information (`c.info` and `s.info`), as shown in the transitions c_3, c_4, c_5 and s_4, s_5, s_6. The control information represent a set of communication parameters that includes: host identifier, transmission checkpoints, and control flags that indicate the role (mobile or stationary node and client or server application) each communicating side plays in a session.

To generate a host identifier, we suggest the use of the SHA-1 algorithm to digest unique RSA Public Keys. The hash SHA-1 allows a low probability of bit collision as well as a huge name space of 2^{160} bits in length. This key material can be used for authenticating peers during handshake. This prevents spoofing and replay attacks during the (re)connections. The key material can be combined with a mutual challenge-question authentication [13], as we implemented in [3].

Thus, the server is aware of who the client is, and vice-versa, and also of where such (re)connection is coming from. Then, the server is able to abort (re)connections that come from untrusted clients, as shown in the transition from s_5 to s_{11}. In the meantime, the client denies communicating with fake servers by aborting connections in the transition from c_5 to c_{11}.

B. Saving the transmission status

After handover, the data enqueued in the send buffer may be lost. The lost data are those unconfirmed by the remote peer and those ready for sending. To provide reliability at the Application Layer, it needs to preserve a copy of the message to be sent, and checkpoints to successfully count the sent and received bytes.

When the peers are connected and ready for sending and receiving messages, the client is in state c_5 and the server is in state s_6. The client sends the message (`send(new.data)`) and reaches state c_6. Then, it saves the transmission context by keeping a local copy of the sent message (from c_6 to c_7) in an ancillary buffer, and it counts the amount of sent bytes (from c_7 to c_8).

After sending, a client can remain waiting for a server's response. When the socket receive buffer is ready for reading, the client consumes the enqueued data and reaches the state c_9 (from c_5 or from c_8). Then, it saves the transmission

 ISBN: 978-1-61208-279-0

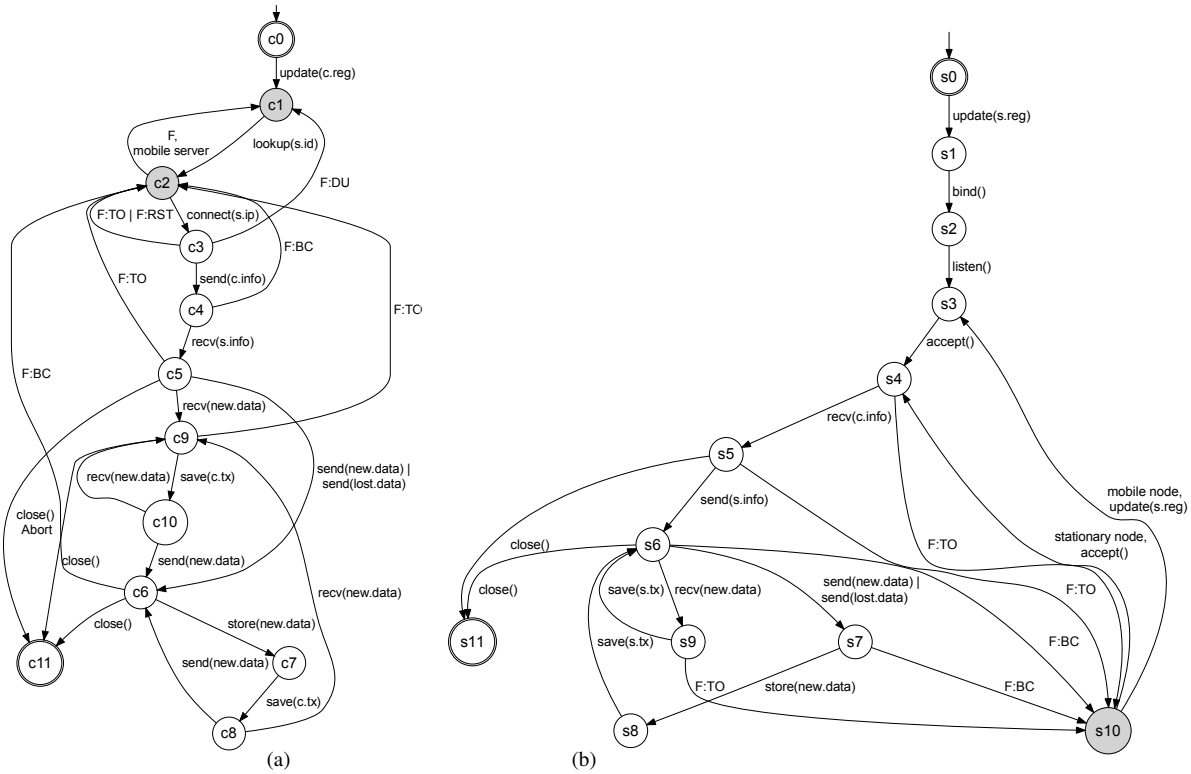


Figure 1. Finite state machines for mobility-aware applications: (a) client, and (b) server.

status by counting the received bytes (c_9 to c_{10}). The client can receive more messages (c_{10} to c_9), send new messages (from c_{10} or from c_8), or finish the transmission (from c_9 or from c_6).

At the server side a similar approach is used. Once connected, the server receives a new message (s_6 to s_9) from the client and saves its transmission context by counting the received bytes (s_9 to s_6). The server can also send new messages to the client (s_6 to s_7) and save its transmission context by copying the sent message in its ancillary buffer and counting the sent bytes (s_7, s_8, s_6). The server is also able to finish the transmission by closing the connection and reaching the final state s_{11}.

C. Disruption Detection

Modifications in the routing table (e.g.,, adding new IPs or routes) and connection timeout break established TCP sessions. On the attempt of sending a new message on a broken TCP socket, an error is returned after calling the send function. In the client, the handling is conducted in the transition from state c_5 to c_6, in which the error `F:BC` (*broken connection*) leads the system to state c_2 again. In the server, the same failure leads to the erroneous state s_{10}.

In the receiver, unread bytes can remain in the socket receive buffer. Thus, the receiver will not crash immediately after the disruption. Instead, the connection becomes half-open, so that the receiver stays in deadlock waiting for messages from a dead sender. To quickly detect a broken connection in the receiver's side, parameters of the TCP Keepalive is reduced as much as possible [14]. Thus, the receiver detects disruption with connection timeouts `F:TO` according to the transitions c_5, c_9, c_2 and s_6, s_9.

D. Resuming communication from broken connections

After a failure, when the transition leads to state c_2, a re-connection has to be established to resume the communication in the client. In order to do so, the server must be resynchronized and prepared to accept a new re-connection from the client again. Then, the server must be waiting in state s_3. By executing the session handshake, client and server are aware about the current transmission context of each other, as shown in the transitions from c_2, c_3, c_4, c_5 and from s_3, s_4, s_5, s_6. This resumes the end-to-end communication into a new TCP connection. The session is preserved by authenticating the peer using the key material and the host identifier. Thus, the server relates the incoming re-connection to an existent session.

E. Restoring Transmission Status

After the TCP connection re-establishment and the handshake between the peers, the sender restores the transmission context by resending the lost data recovered from the ancillary buffer. To determine the amount of lost data, both client and server must know the amount of bytes that was sent and

received by each other until to the disruption. To do so, the peers can mark transmission checkpoints by counting the bytes that are sent and received successfully.

We suggest the use of in-band signalling between the peers. Session control information, including transmission checkpoints, are exchanged over the data connection, as we propose with handshake between client and server with transitions from c_3, c_4, c_5 and from s_4, s_5, s_6, respectively.

The sender's amount of data lost in handover is given by:

$$N_{ld} = N_{ls} - N_{rr}, \quad (2)$$

where N_{ls} is the local amount of bytes sent by the node, and N_{rr} is the amount of bytes received by the remote peer.

The node is able to recover the lost data from the ancillary buffer and, then, resend them to the opposite node with the function `resend(lost.data)` in the transition from c_5 to c_6 and from s_6 to s_7.

While resending, however, a disruption might eventually cause a connection failure again. Since the lost data are already saved, only new messages are copied into the ancillary buffer. The model provides the same support for handling this disruption as described previously. In that case, the remote peer can receive part of the lost data that the sender tried to resend. Although this sending is incomplete, with the next handshake the peers recalculate the new amount of lost data, which denotes the remaining of lost data that has to be sent. This strategy provides integrity of the data transmission in the whole session duration and brings reliability to the Application Layer of the mobile nodes.

F. Mobility scenarios and Location Management

In a *Single Jump* scenario, the server usually runs on a stationary host to which the client re-connects. The mobility becomes a more complex problem when the server node also changes its location. In this scenario, called *Double Jump*, location management is required. Some entity must provide the current location of mobile server nodes. Besides, there is the possibility of moving both nodes simultaneously, which increases the chance of race conditions.

Race conditions are experienced when the client tries to re-connect to the server when it is not prepared to accept connections. This occurs when: i) the server visits a new network and the client does not know about the new server's location; ii) the client detects the disruption before the server, so that the client starts the re-connection in advance.

Mobility protocols can leverage Dynamic DNS for mobile node location [7]. *Rendezvous Servers* [8][10] have also been applied to location management. However, we suggest the use of distributed mechanisms to store node locations. A Distributed Hash Table (DHT) of general purpose, such as OpenDHT [15], is able to provide the necessary scalability to manage mobile node location. A DHT provides storage based on key-value semantic, as well as a simple put-get interface. Thus, to query the peer's location, a node uses the peer's host identifier as a key for searching in the DHT. To update location, a node puts its location register under its host identifier in the DHT.

A node is aware of the role it plays in a session. Thus, when the server is migrated, if it runs on a mobile node, it puts its location register in the DHT (s_{10}, s_3). Otherwise, it just waits for re-connections from the client (s_{10}, s_4), without performing a location update. In the client, a connection failure arises when the server is migrated. The client is aware of the role the server plays due to the control flags exchanged in the handshake. Then it can query the current server's location in the DHT with the server's host identifier. If the client attempts to re-connect to the outdated server's location, the connection fails, since the destination is unreachable (`F:DU`). Then, the client stays in the loop from c_2, c_3, c_1 until the race condition is overcome.

III. HANDOVER DELAY MODEL

To evaluate mobility-aware applications we applied the mathematical model proposed by Mohanty and Akyildiz [11] and extended by Shah *et al.* [12].

As discussed in [12], the end-to-end handover delay is the sum of the following delays:

$$T_{handover} = T_{L2} + T_{IP} + T_{sig} + T_d, \quad (3)$$

where they represent, respectively, the delay of: Link Layer connectivity, IP address acquisition, handover control messages signalling, and one-way data packet transmission in new network.

The critical delay is T_{sig}. It varies according to the approach used by the mobility protocol. Next, we discuss the model used in [11] and [12], which is based on the number of control messages exchanged between mobile node (MN) and correspondent node (CN) to complete end-to-end handover.

A. Handover signalling delay as function of packet loss probability

The handover operation model of TCP-Migrate [7] discussed by [11] can be applied to the majority of end-to-end mobility management protocols [12]. As in [16], the model is based on the delay of TCP connection establishment as function of end-to-end packet loss probability and of retransmission timeout. In mobility scenarios, then, the model represents the delay of connection set-up or migration between MN and CN in new visited network. It is calculated as function of end-to-end packet loss probability, number of handover signalling messages exchanged between the peers and their number of retries, and the retransmission timeout.

The delays of handover signalling completion for end-to-end mobility protocols of two [12] and three control messages [11], respectively, are

$$L_{sig}(i,j) = RTT + (2^i + 2^j - 2)RTO, \quad (4)$$

 ISBN: 978-1-61208-279-0

and,

$$L_{sig}(i,j,k) = 1.5RTT + (2^i + 2^j + 2^k - 3)RTO, \quad (5)$$

where i, j, and k are the number of unsuccessful tries for handover control messages, RTT is the round trip time of the message, and RTO is the initial retransmission timeout for the TCP connection.

Using Equation (4) and (5), we express the overall delay of handover signalling completion for a end-to-end mobility protocol of N_m handover control messages as

$$L_{sig}(A) = \frac{N_m}{2} RTT + \left[\left(\sum_{i=1}^{N_m} 2^{a_i} \right) - N_m \right] RTO, \quad (6)$$

where $A = (a_1 \ a_2 \ a_3 \ \ldots \ a_{N_m})$ is the number of unsuccessful tries for each handover control message.

The end-to-end packet loss probability between MN and CN discussed by [11] is

$$p = 1 - (1 - FER)^f (1 - p_{wr}), \quad (7)$$

where FER is the link layer frame error rate; $f = \frac{L_p}{L_f}$ is the number of link layer frames per packet, where L_p is the length of the packet, and L_f is the length of the Link Layer frame; and p_{wr} is the packet loss probability in wired link part.

From the TCP latency model in [16], the authors in [12] and [11] discuss the probabilities of handover signalling completion for protocols of two and three control messages, respectively, by expressing

$$P_{sig}(i,j) = p_1^i (1-p_1)\ p_2^j (1-p_2), \quad (8)$$
$$P_{sig}(i,j,k) = p_1^i (1-p_1)\ p_2^j (1-p_2)\ p_3^k (1-p_3), \quad (9)$$

where p_1, p_2, and p_3 are the packet loss probabilities calculated using Equation (7) for each individual handover control message with its respective i, j, and k number of unsuccessful tries. These equations represent the probabilities of end-to-end mobility protocols to complete handover after the exchange of i unsuccessful tries for the first control message, followed by the first succeeded message, followed by j unsuccessful tries for the second control message, followed by the second succeeded message, and so on.

Assuming the number of unsuccessful retries in A, for a mobility protocol of N_m control messages, we express the overall probability of handover signalling completion as

$$P_{sig}(A) = \prod_{i=1}^{N_m} p_i^{a_i} (1 - p_i), \quad (10)$$

where p_i is the end-to-end packet loss probability calculated using Equation (7) of ith control message with its respective a_i number of unsuccessful tries.

The average of the handover signalling messages delay for a protocol of $N_m = 3$ control messages, such as TCP-Migrate, is [11]:

$$E[T_{sig}] = \sum_{i=0}^{N_r-1} \sum_{j=0}^{N_r-1} \sum_{k=0}^{N_r-1} P_{sig}(i,j,k)\ L_{sig}(i,j,k), \quad (11)$$

where N_r is the number of retransmissions before giving up and abort the connection establishment of TCP-Migrate. As discussed in [12], N_r is used as the maximum allowed number of retries for handover control messages.

Using the overall delay and probability of Equation (6) and (10), we generalize the average of the handover signalling messages delay for any N_m-message protocol as

$$E[T_{sig}] = \sum_{A=1}^{N_A} P_{sig}(A_i)\ L_{sig}(A_i), \quad (12)$$

where N_A is the number of all the combination of A taking into account N_m and $N_r - 1$.

B. Handover signalling delay for mobility-aware applications

According to the state machines for mobility-aware application, the handover signalling delay at the active opener (mobile client) in single jump scenario is

$$T_{sig_{SJ}} = T_{TCP} + T_{hs} + T_{ld}, \quad (13)$$

where T_{TCP} is the new TCP connection establishment delay between MN and CN (from c_2 to c_3), T_{hs} is the protocol handshake delay which is used for end-to-end synchronization and/or authentication (from c_3 to c_5), and T_{ld} is the lost data retransmission delay (from c_5 to c_6). These delays are calculated using Equation (12), each one according to the number of control messages N_m exchanged between MN and CN.

The handover signalling delay at the active opener in double jump scenario is

$$T_{sig_{DJ}} = T_{lookup} + T_{TCP} + T_{hs} + T_{ld}, \quad (14)$$

where T_{lookup} is the client's delay to lookup server's location. T_{TCP}, T_{hs}, and T_{ld} are defined in Equation (13).

To lookup location register, the requester sends a request to the DHT and waits response from it. The DHT is accessed using either Sun RPC over TCP or XML RPC over HTTP [15]. Assuming that the delay to get a register is the sum of the time to establish a TCP connection with the DHT and the message round trip time (lookup request and its reply), the registration location lookup delay can be expressed as $T_{lookup} = T_{TCP} + T_{req} + T_{rep}$.

 ISBN: 978-1-61208-279-0

However, as discussed in [11], the round trip time for a lookup request can be expressed as twice the one-way end-to-end delay transportation. Thus,

$$T_{lookup} \quad = \quad T_{TCP} + 2(D + t_w), \qquad (15)$$

where D is the Link Layer access delay, and t_w is the one-way delay in the wired network between the new Base-Station/Access-Point and the remote node (DHT).

IV. Simulation Results

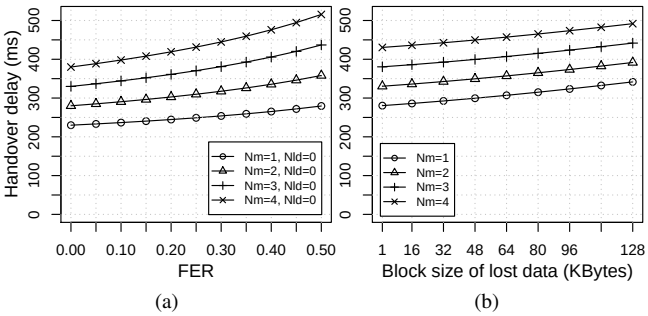
We evaluated the cost of handovers for the mobile active opener by using the described handover delay model in both single and double jumps; i.e., the handover latencies for the client to leave the state c_1 (in double jump) or c_2 (in single jump) and, then, reach the state c_5 (if there are no lost data to resend) or reach the state c_9 (after resending the lost data). Both communicating sides accomplish equivalent operations to handle F, therefore, the cost of handover is considered the same for both client and server. We implemented the handover delay model using the GNU software environment for statistical computing provided by R[17].

A. Simulation Parameters

In most application protocols the connection is considered to be established when SYN/ACK packet arrives at the active opener [16]. This is because immediately after sending ACK in the third packet of the three-way handshake, the active opener sends a data segment to the passive opener that contains the redundant ACK [16]. Then, we assume that TCP connection is considered to be established by the client with 2 control messages. Thus, we set $N_m = 2$ in Equation (12) to calculate T_{TCP}.

The number of control messages exchanged during the handover signalling depends on the synchronization and/or authentication used by the mobility protocol. The mobile node can confirm its new location by simply sending a single control message to the correspondent node. SIP [8] requires at least exchanging two control messages using re-INVITE and 200 OK confirmation. The FSM suggests exchanging two control messages to provide end-to-end synchronization and unilateral authentication. TCP-Migrate [7] requires exchanging three control packets with the three-way handshake needed to migrate TCP connections. When secure connection re-establishment is a requirement, a four-way handshake can provide mutual challenge-response authentication and end-to-end synchronization [18] [3]. We evaluated these different handover signalling strategies for mobility-aware application by varying the values of N_m from 1 to 4 in the calculation of T_{hs} from Equation (12).

To compare these different handover signalling strategies, we assume the values of the handover delay model parameters as in [12]: $T_{L2} = 10$ ms, $T_{IP} = 20$ ms, $T_d = 50$ ms, $RTT = 100$ ms, $RTO = 200$ ms, $pwr = 1e-6$. Since most TCP implementations abort connection establishment attempts after 4-6 failures [16], we assume $N_r = 4$ retries. We set $L_f = 576$ bytes (length of Link Layer frame), regarding the packet size every host must be able to handle on the Internet [19]. The length of the packet $L_p = 40$ bytes for the TCP connection establishment, regarding the size of IP header and TCP header, both with no options. We assume $L_p = 60$ bytes for the length of the handover control message, which is able to accommodate the control information (host identifier, checkpoints, and protocol flags) we described in Section II-A. As in [11], we set the link layer access delay $D = 10$ ms in WLANs, and the one-way delay in the wired network $t_w = 50$ ms.



(a) (b)

Figure 2. Handover delay in single jump scenarios: (a) as function of the frame error rate; (b) as function of the lost data (N_{ld}) in handover.

As argued in [12], although these values are not constant and may vary depending on the network conditions, when the values are changed they affect in the same manner the behaviour of all the mobility strategies evaluated.

B. Handover Latencies

Figures 2 and 3 show handover delay results for $N_m = 1, 2, 3, 4$ strategies in single and double jumps, respectively. Figures 2(a) and 3(a) show delays as function of FER when there are no lost data in handovers, i.e., $N_{ld} = 0$, hence, $T_{ld} = 0.0$. In both scenarios, handover delays increase as the FER increases, which is natural. When FER gets high, it increases the probability of dropping packets and, therefore, handover control messages need to be retransmitted N_r times to successful complete the handover. The overall cost of using a handover control message was 62.18 ms in average for both mobility scenarios.

Figures 2(b) and 3(b) show delays as function of the lost data in handovers. We set FER $= 1e-3$ and evaluated the lost data by varying N_{ld} in L_p with multiple blocks of 16 KB up to 128 KB, which is the maximum amount of lost data by a sender. We assume such limit due to the maximum size of 131071 bytes for socket send buffers in GNU/Linux systems with Kernel version above 2.4. We observed that the cost of retransmitting a block of 16 KB of lost data with a single operation `send(lost.data)` increases the handover latency in 7.65 ms.

 ISBN: 978-1-61208-279-0

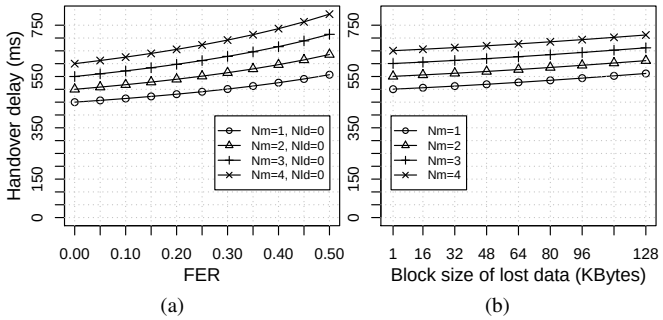


Figure 3. Handover delay in double jump scenarios: (a) as function of the frame error rate; (b) as function of the lost data (N_{ld}) in handover.

In single jumps, which is the common mobility scenario, the mobile client can be moved with no need for location updates. Therefore, the entity for managing location is avoided. This allows reduced handover latencies than the ones in double jumps, as shows Figure 3. Due to the registration location lookup delay (T_{lookup}), in which the client gets the current server's location registration from the DHT in order to re-connect, the handover latencies in double jumps are 42% in average bigger than the ones in single jump; i.e., a mean overhead of 244.37 ms.

We observed that mobility-aware applications provide little and smooth impact when both FER and lost data increase. This behaviour can be an interesting advantage in dynamic scenarios, such VCNs, where devices are of high mobility and packet loss rates are prone to be higher than the ones in mobility scenarios focused on the mobile user.

Comparing results in [12], $N_m = 2$ strategy (as primarily suggest our FSMs) provides handover latencies very close to the kernel space based mobility solutions, such as TCP-R, TCP-Migrate and HIP. In relation to these solutions, an overhead lower than 100 ms is the weight of the burden on handling mobility at the user address space of mobile nodes. However, this performance degradation is derisive when taking into account the benefits of providing easy deployment and maintenance of full support for mobility, which can be implemented as a single software service.

V. Conclusion

In this paper, we described the general purpose mobility-aware applications by means of Finite State Machines capable of providing transitions between erroneous and safe states. These transitions are abstract operations based on the Socket semantic, which are necessary to resume consistently and lossless the end-to-end communications broken with mobility. The presented abstraction is useful in support for developing mobility solutions as software services at the end nodes involved in TCP sessions, with no relying on home networks or agents to intermediate the communication. Applying the analytical model for handover delay discussed in [11] and [12], simulation results show good performance and smooth impact in situations of high packet loss rates and of lost data.

VI. Acknowledgements

We thank INCT-SEC for funding this work by means of the agencies CNPq and FAPESP.

References

[1] C. E. Perkins, "IP Mobility Support for IPv4, Revised," in *IETF RFC 5944*, November 2010, pp. 1–100.

[2] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6," in *IETF RFC 6275*, July 2011, pp. 1–169.

[3] B. Y. L. Kimura, H. C. Guardia, and E. S. Moreira, "Disruption-Tolerant Session for Seamless Mobility," in *WCNC 2012: Proceeding of the 2012 IEEE Wireless Communications and Networking Conference*, 2012, pp. 2412 –2417.

[4] B. Y. L. Kimura and H. C. Guardia, "TIPS: Wrapping the Sockets API for Seamless IP Mobility," in *SAC'08: Proceedings of the 23rd Annual ACM symposium on Applied computing*, vol. 3, March 2008, pp. 1940–1945.

[5] W. M. Eddy, "At what layer does mobility belong?" *IEEE Communications Magazine*, vol. 42, no. 10, pp. 155–159, October 2004.

[6] D. Funato, K. Yasuda, and H. Tokuda, "TCP-R: TCP mobility support for continuous operation," in *ICNP'97: Proceedings of the IEEE International Conference on Network Protocols*, October 1997, pp. 229 – 236.

[7] A. C. Snoeren and H. Balakrishnan, "An end-to-end approach to host mobility," in *MobiCom'00: Proceedings of the 6th ACM Annual International Conference on Mobile computing and Networking*, August 2000, pp. 155–166.

[8] J. Rosenberg, et al., "SIP: Session Initiation Protocol," in *IETF RFC 3261*, June 2002, pp. 1 – 269.

[9] V. C. Zandy and B. P. Miller, "Reliable Network Connections," in *MobiCom'02: Proceedings of the 8th ACM Annual International Conference on Mobile computing and Networking*, September 2002, pp. 95–106.

[10] R. Moskowitz, P. Nicander, and P. Jokela, "Host Identity Protocol," in *IETF RFC 5201*, April 2008, pp. 1 – 104.

[11] S. Mohanty and I. F. Akyildiz, "Performance analysis of handoff techniques based on mobile ip, tcp-migrate, and sip," *IEEE Transactions on Mobile Computing*, vol. 6, no. 7, pp. 731–747, July 2007.

[12] P. A. Shah, M. Yousaf, A. Qayyum, and H. B. Hasbullah, "Performance comparison of end-to-end mobility management protocols for tcp," *Journal of Network and Computer Applications*, vol. 35, no. 6, pp. 1657 – 1673, 2012.

[13] D. M'Raihi, J. Rydell, S. Bajaj, S. Machani, and D. Naccache, "OCRA: OATH Challenge-Response Algorithm," in *IETF RFC 6287*, June 2011, pp. 1 – 38.

[14] B. Y. L. Kimura, R. S. Yokoyama, R. R. F. Lopes, H. C. Guardia, and E. S. Moreira, "Prototyping applications to handle connection disruptions in end-to-end host mobility," in *WONS 2010: the Seventh IEEE International Conference on Wireless On-demand Network Systems and Services*, 2010, pp. 1–8.

[15] S. Rhea, B. Godfrey, B. Karp, J. Kubiatowicz, S. Ratnasamy, S. Shenker, I. Stoica, and H. Yu, "OpenDHT: a public DHT service and its uses," *SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 4, pp. 73–84, August 2005.

[16] N. Cardwell, S. Savage, and T. Anderson, "Modeling TCP latency," in *INFOCOM 2000: Proceedings of the Nineteenth IEEE Annual Joint Conference of the IEEE Computer and Communications Societies*, March 2000, pp. 1742 – 1751.

[17] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008. [Online]. Available: http://www.R-project.org

[18] B. Y. Kimura, R. S. Yokoyama, H. C. Guardia, and E. S. Moreira, "Secure connection re-establishment for session-based ip mobility," in *CBSEC 2012: Proceeding of the 2nd Brazilian Conference on Critical Embedded Systems*, 2012, pp. 58–63.

[19] J. Postel, "Internet Protocol," in *IETF RFC 791*, September 1981, pp. 1 – 45.

 ISBN: 978-1-61208-279-0

Dynamic IMS Reconfiguration using Session Migration for Power Saving

Satoshi Komorita, Manabu Ito, Yoshinori Kitatsuji, and Hidetoshi Yokota
KDDI R&D Laboratories, Inc.
Fujimino-shi, Saitama, JAPAN
e-mail: {sa-komorita, mn-itou, kitaji, and yokota}@kddilabs.jp

***Abstract*— Legacy telecommunication services have been tending to shift to All-IP networks from closed and circuit switched networks by the deployment of high bandwidth and stable mobile broadband networks such as Long Term Evolution (LTE). The shift to an All-IP network leads to the integration of telecommunication services and internet services, which are called Over-The-Top (OTT) services. The trend towards increasing and unpredictable traffic makes it important for the IP Multimedia Subsystem (IMS), which is a call control system running over the All-IP network, to be flexible and reliable. With that goal, the authors have proposed an IMS reconfiguration mechanism using call session state migration, which allows the IMS to distribute live call sessions as required by the failure of IMS servers and the quantity of calls in progress. Furthermore, in this paper we propose a model to determine the configuration of the IMS dynamically from the perspective of electrical power consumption and performance. A formula based on the model calculates which IMS servers should be active and which call sessions should move to which IMS servers to reduce the overall power consumption, using linear programming. Moreover, this paper implements the proposal and shows its results.**

Keywords-IMS; SIP; Session Migration; Power Saving; Linear Programming; Optimization; CPLEX

I. INTRODUCTION

In order to provide voice over IP (VoIP) service, mobile network operators (MNOs) are developing the IP multimedia subsystem (IMS) [1], which is a call control system over an IP network (adopted in GSMA [2]), to ensure interconnectivity between MNOs around the world. MNOs also plan to launch the Rich Communication Suite (RCS) [3], which will enable mobile phone users to use instant messaging, live video sharing, and file transfer over the IMS. In addition, through leveraging their IMS development, MNOs are now looking into the possibility of integrating Over-The-Top (OTT) services such as Google services, Facebook, and Skype.

The support of these future services in addition to legacy telecommunication services such as voice and Short Message Service (SMS) has resulted in an increase in signaling traffic in the IMS. It is difficult to estimate the required system capacity for the increase because OTT services are created by third parties outside MNO's awareness. To continue to deal with increasing and unpredictable traffic without investing in extra capacity, enhanced flexibility is desirable in the operation of the call session control functions (CSCFs) of which the IMS is composed. For example, in case of server failure or disaster, it is desirable to ensure reliability by continuing the processing of CSCFs on other servers, thus preventing service disruption. On the other hand, it is desirable to minimize the number of running servers, to decrease power consumption.

In order to realize the desired flexibility, we have proposed call session state migration [4], which allows call/service sessions to continue to be processed on another CSCF. The proposed method mainly involves a migration of the session states between CSCFs and a routing mechanism for control messages in the IMS without extensions to the standard procedure for control of a call/service session. However, the MNOs still need to decide which sessions should move to which CSCFs and which server can be stopped, even if the method provides for flexibility in the IMS configuration. This makes IMS reconfiguration more difficult as the number of CSCFs increases.

Thus, in this paper, we propose a model of IMS reconfiguration to determine the IMS configuration dynamically and easily so as to minimize power consumption and provide enough performance. The model incorporates the power consumption of the IMS so that the IMS can provide enough performance, while considering the power consumption of the reconfiguration, including all session state migration and running CSCFs. We solve the model and determine the reconfiguration by using linear programming [5]. Furthermore, we implement our proposal and the measurement metrics required for our model, and show the result of our proposal in a real environment.

The rest of this paper is organized as follows. Section 2 describes the IMS architecture and the call session state migration. Section 3 proposes a model of the IMS reconfiguration and its formula for linear programming. Section 4 shows our implementation and results of the experiment. Finally, Section 5 concludes this paper.

II. CALL SESSION STATE MIGRATION

A. IMS Architecture and Call Session Control

Fig. 1 shows the basic IMS network configuration. The following IMS components are located in the IMS core network that an MNO manages and operates: an Home Subscriber Server (HSS), which is a database server for managing subscribers, an Serving Call Session Control Function (S-CSCF), which is the main Session Initiation Protocol (SIP) [6] server for call control, a Proxy CSCF(P-

 ISBN: 978-1-61208-279-0

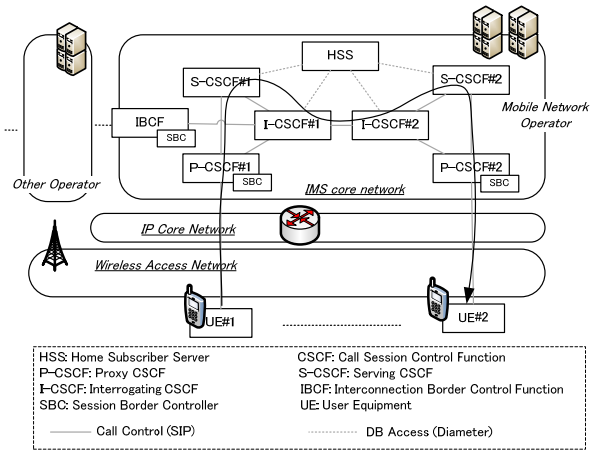


Figure 1. Basic IMS Architecture

CSCF) that communicates with the User Equipment (UE) directly and establishes a secure connection to it, an Interrogating CSCF (I-CSCF), which is a kind of resolver of SIP message routing, and an Interconnection Border Control Function (IBCF) that is a gateway for other MNO's IMSs. Furthermore, an Session Border Controller (SBC) providing a variety of functions for security and connectivity (e.g., access control, topology hiding, NAT traversal, protocol interworking, and media monitoring) is defined and often integrated into the P-CSCF and the IBCF in a real operation. A UE connects with the IMS through the IP core network and wireless access networks such as EV-DO [7] and Long Term Evolution (LTE) [8].

In the IMS, SIP is used for call control between CSCFs and UEs. First, a UE registers with the IMS before obtaining IMS services. The UE conducts its registration with an S-CSCF via a P-CSCF and an I-CSCF. The S-CSCF assigned by the I-CSCF verifies the UE based on its authentication information stored in the HSS. After that, the UE uses IMS services by sending and receiving SIP messages to/from a correspondent UE via the P-CSCF, S-CSCF, and I-CSCF. For example, when the UE makes a call, the UE sends an INVITE message to the correspondent UE and establishes a session to communicate the required information. The exchanged SIP messages take, as shown in Fig. 1, the following path: UE#1 (caller), P-CSCF#1, S-CSCF#1, I-CSCF#1, I-CSCF#2, S-CSCF#2, P-CSCF#2, and UE#2 (callee). Note that the assigned P-CSCF and S-CSCF of UEs hold states of register and sessions and each SIP message includes route information to their CSCFs, thus once CSCFs begin to handle the session of the UEs, it is not possible to change CSCFs to process the sessions in the standard procedures.

B. Call Session State Migration and Determinaton of IMS Reconfiguration

We have proposed a call session state migration mechanism [4], which allows sessions to continue to be processed on another CSCF in order to cope with server failure and dynamically change the number of physical servers to match the network and resource usage situation.

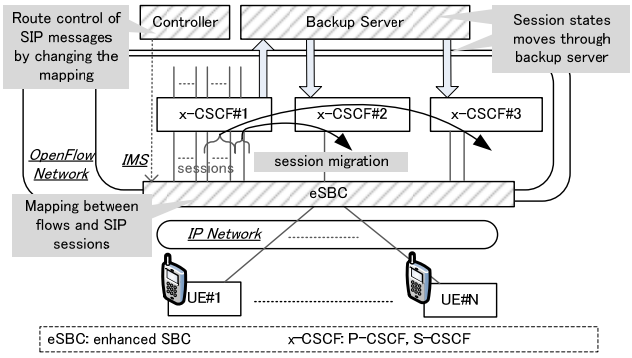


Figure 2. Call State Session Migration

The method mainly consists of the migration of session states between CSCFs and route control of SIP messages as shown in Fig. 2. The session states are migrated through a backup server in order to backstop server failure. The routing of SIP messages is controlled by using OpenFlow [9] to hide changes in the IP address of CSCFs. OpenFlow is a flow-based routing protocol, and we enhanced SBC to map between SIP sessions and flows of the OpenFlow network. This mechanism supports granular session migration to realize effective resource use.

While the session migration mechanism has been proposed, MNOs still need to decide which sessions should move to which CSCFs and which server can be stopped. If there is obvious trouble such as a server failure, MNOs can decide easily. However, if there are a few hundred servers in the IMS, it is difficult to estimate outcomes and quickly make a decision appropriate to the situation. Thus, a model to determine the IMS configuration dynamically is needed to support the operation of a flexible IMS.

C. Related Work

The study of optimized configuration in networks and SIP servers has a long history of improving performance such as traffic throughput, and maximizing the amount of processing using limited resources. Some studies [10][11] tried load balancing of processing to improve the performance of SIP servers. Another investigation [12] tried to propose an effective route for reducing message delay. Analysis of the IMS, which has additional processing compared with basic SIP procedures, has also been undertaken. N. Rajagopal *et al.*[13] analyzed IMS processing based on queuing of SIP messages and estimated the overload of each CSCF based on the SIP processing sequence.

The above works focused mainly on calculating the capability of the IMS in advanced before operation. On the other hand, the concept of effective resource use on demand has been also proposed, based on the perspective of flexible resource balancing using the recent development of virtualization technologies, rather than improvement in performance using limited resources [14]. This study proposed live session migration between physical servers by using a virtualized machine's live migration on demand. V. Petrucci *et al.* [15] have proposed the concept of calculating

 ISBN: 978-1-61208-279-0

the required minimum number of servers and stopping the extra servers to save power in cluster environments.

However, in dynamic IMS reconfiguration, we need to consider the specific IMS configuration and the cost of the reconfiguration, and in particular, live session migration. In addition, some metrics for the calculation must be measured in a real environment to evaluate their adequacy.

III. PROPOSED MODEL OF DYNAMIC IMS RECONFIGURATION

A. Modeling Aproach based on Power Consumption

As an index to characterize IMS reconfiguration, we focus on the reduction of power consumption of the IMS because the purpose of a flexible IMS is to supply sufficient IMS capacity on demand while using minimum facilities. In our approach, we estimate all power consumption required for running the IMS and executing the reconfiguration, and use a linear programming method to try to reduce the total amount while still providing stable IMS service.

In contrast to a static IMS configuration, a dynamically reconfigurable IMS can change its configuration as many times as needed depending on the momentary situation, such as the number of calls. Thus, we also consider the cycle of IMS reconfiguration. Otherwise, even if the number of servers and their power consumption could be reduced on a temporary basis, the total power consumption might actually increase if the reconfiguration and call session state migration happened frequently.

B. Reference Model of IMS Reconfiguration

Fig. 3 shows a reference model of IMS reconfiguration, consisting of two IMS configurations and the transition between them. In the IMS, there are several nodes, which are fundamental physical servers. An Functional Entity (FE) is a running and session-processing entity such as a P-CSCF on a node. Here, for simplicity, just one FE will run on one node and just one type of FE shall exist in this model so that we can consider this model for each type of FE independently. In addition, nodes in the IMS can be turned off and on for power saving.

The following parameters are introduced for nodes and FE of the reference model.

$$
\begin{aligned}
&N = \text{a set with all nodes}\\
&B_n = \text{base power consumption by node\#}n,\ n \in N\\
&y_n = \begin{cases} 1 \text{ if node\#}n \text{ is on} \\ 0 \text{ if node\#}n \text{ is off} \end{cases}, \{y_n\} \text{ binary (integer)}\\
&M_n = \text{total number of sessions on node\#n}\\
&K_n = \text{capacity of FE on node\#}n\\
&\tau = \text{base power consumption of FE}\\
&\varepsilon = \text{power consumption for running a single session}\\
&Z \equiv \text{a state of configuration given by } \{y_n, M_n\}, n \in N\\
&R = \text{runing time of state Z}
\end{aligned}
\tag{1}
$$

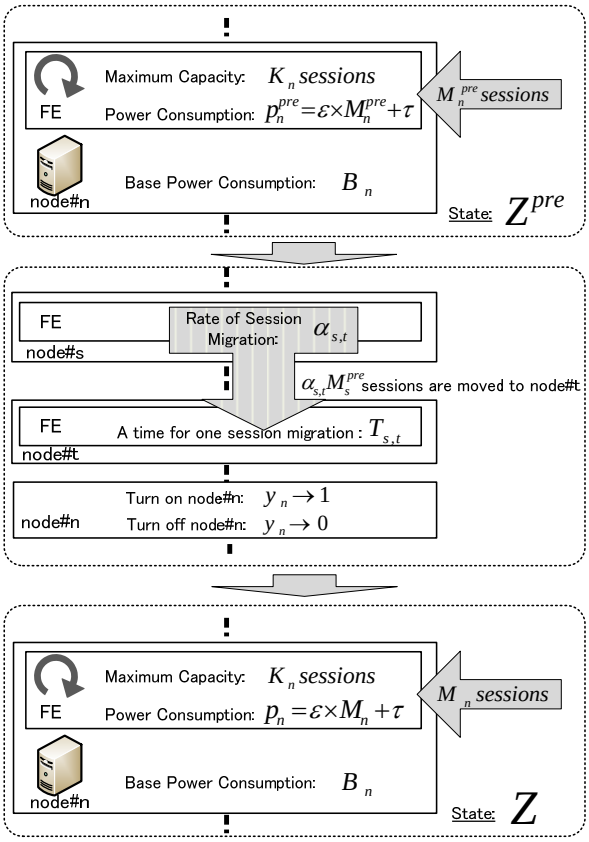


Figure 3. Reference Model of IMS Reconfiguration

Then, the IMS executes the reconfiguration by performing session migrations and turning nodes on/off. Previous state Z^{pre} makes a transition to the next state Z. The parameters for the transition are as follows:

$$
\begin{aligned}
&\alpha_{s,t} \equiv \text{fraction of sessions moved from FE on node\#}s\\
&\qquad \text{to FE on node\#}t, s \in N, t \in N\\
&C_{s,t} \equiv \text{power consumption of moving a single session}\\
&\qquad \text{from FE on node\#s to node\#t. Note that } C_{n,n} = 0\\
&T_{s,t} \equiv \text{time for moving a single session}\\
&\qquad \text{from FE on node\#}s \text{ to node\#t}\\
&\theta_n = \left| y_n - y_n^{pre} \right|,\ y_n^{pre} = \text{previous on/off state of node\#}n\\
&h_n = \text{power consumption of switching state at node\#}n
\end{aligned}
\tag{2}
$$

After the reconfiguration, the number of sessions can be calculated as below.

$$M_n = \sum_{s \in N} \alpha_{s,n} M_s^{pre} \tag{3}$$

In addition, several constraints can be considered for the proposed reference model as below. Here, we defined thresholds:

$U \equiv$ maximum utilization rate of FE
$W \equiv$ maximum authorized time for reconfiguration (4)

C. Formulation for Linear Programming

To obtain the most suitable reconfiguration by using linear programming, in this section we will formulate the reference model of the previous section. Here, the objective function to be minimized by linear programming is the power consumption required for running the IMS and executing the reconfiguration.

The power consumption of the reference mode falls into three categories. The first is running power, which is consumed by running nodes and processing sessions. It depends on their running time, thus it is given by:

$$PCR = R\sum_{n\in N} y_n \left[B_n + \varepsilon \times M_n + \tau\right] \quad (5)$$

Secondly, the power is consumed by switching node states such as booting and shutdown. The total power consumed for them is given by:

$$PCS = \sum_{n\in N} h_n \theta_n \quad (6)$$

Thirdly, the migration of sessions from one FE to another consumes power in signaling overhead. The total power consumed for migrating sessions between nodes is given by:

$$PCM = \sum_{s\in N}\sum_{t\in N} C_{s,t}\alpha_{s,t} M_s^{pre} \quad (7)$$

Then, the objective function to be minimized can be expressed as the sum of the above:

$$OF = PCR + PCS + PCM \quad (8)$$

In addition to the objective function, there are several constraints for the linear programming to take into account. The number of sessions cannot be larger than the FE capacity and its utilization to provide stable IMS capacity. The constraint is given by:

$$M_n \le K_n U \quad (9)$$

The session transfer time is assumed to be lower than a given threshold in order to satisfy quality-of-service requirements because a large reconfiguration has an impact on stability of the IMS service. The constraint to suppress excessively large reconfigurations is given by:

$$\sum_{s\in N}\sum_{t\in N} T_{s,t}\alpha_{s,t} M_s^{pre} \le W \quad (10)$$

Finally, the ratios of sessions to be migrated and to be kept on the same node should sum to one. This constraint is given by:

$$\sum_{s\in N}\sum_{t\in N} \alpha_{s,t} = 1 \quad (11)$$

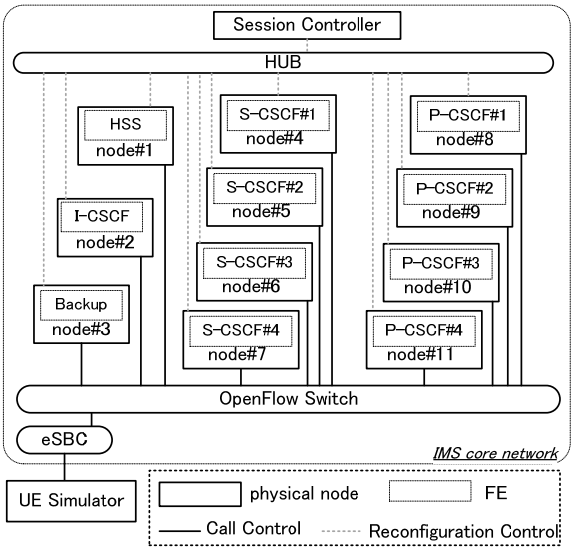


Figure 4. Experimental Network Configuration

D. Cycle of IMS Reconfiguration

In order to reduce the power consumption of the reference model in a certain period of time, the cycle of the reconfiguration is also important because the IMS can execute the reconfiguration as many times as needed in the period. The objective function (8) expresses one reconfiguration from state Z^{pre} to state Z and the current state Z during R. Thus, the total objective function during a certain period which includes state $Z \in Z_{total}$ is as described below.

$$OF_{total} = \sum_{Z\in Z_{total}} OF \quad (12)$$

Z_{total} is a set of state Z during the period. However, it is difficult to solve OF_{total} all at once. Thus, the system evaluates the objective function (8) periodically and executes a reconfiguration based on the result step by step. If the following equation (13) is satisfied, the reconfiguration should be executed. The IMS also executes the reconfiguration if constraint (9) is not satisfied in the current state due to changing of M_n.

$$OF \le R\sum_{n\in N} y_n^{pre}\left[B_n + \varepsilon \times M_n^{pre} + \tau\right] \quad (13)$$

This means that if the power consumption of the reconfiguration (6) + (7) is smaller than the expected reduction going from the previous state Z^{pre} to the next state Z during R, the reconfiguration should be done. However, note that this depends on value of the expected R. Thus, to improve the accuracy of the model, an estimation of R based on the call rate in an actual environment is desirable.

IV. IMPLEMENTATION AND EVALUATION

A. Implementation and Experiment Configuration

To solve our objective function describing the reference model, we used CPLEX [5], which can solve large linear problems. In addition, we implemented the IMS on real

 ISBN: 978-1-61208-279-0

TABLE I. SPECIFICATIONS OF EXPERIMENTAL NETWORK COMPONENTS

Node	Specifications			
	Hardware	CPU	Memory	OS
P-CSCF, I-CSCF, S-CSCF, eSBC, UE Simulator	Generic PC	Core2Duo E8600 3.33GHz	4 GByte	Ubuntu 10.04 LTS
HSS	Generic PC	Xeon E5450 3GHz	12 GByte	
Session Controller	Generic PC	Core2Duo E8400 3GHz	8 GByte	
OpenFlow Switch	NEC UNIVERGE PF5240 (OpenFlow Ver.1.0.0)			

machines to obtain the parameters required for the reference model and to verify our proposal.

Fig. 4 shows our experimental network configuration. In the IMS network, CSCFs and an eSBC are connected via an OpenFlow switch. A UE Simulator connects to the IMS through the eSBC. A session controller connects to the CSCFs, OpenFlow switch, and eSBC via a HUB. An HSS also connects to the CSCFs via the OpenFlow switch. Table. I shows the specifications of the components. The P/I/S-CSCFs and HSS were built based on the Open IMS core [16], which is an open-source SIP server. We implemented the required modules of the CSCFs for our proposed mechanism and the required software for the session controller and eSBC. We simulated a large UE load by using ims_bench [17], which is load testing software for the IMS (also open-source). The controller periodically collects the number of sessions from each CSCF. Then on the basis of its information, the controller calculates the objective function and executes the reconfiguration if the result satisfies the criterion in (13).

B. Measurements

We executed the basic calls and their session migration to obtain the parameters required for evaluating the objective function. UEs registered with four S-CSCFs through four P-CSCFs, and then the UEs continued to make calls to keep certain sessions at each P-CSCF. Here the duration of each call was 120 seconds. Then, we performed call session migration between P-CSCFs and turned the nodes on/off. The result is shown in Table II. The values are averages for four P-CSCFs. Note that here we adopted a P-CSCF as the target FE and measured the parameters. Of course, we can measure S-CSCF in the same way.

On basis of the obtained parameters, we wrote scripts for CPLEX for the objective function and the constraint conditions. Then we verified the behavior of the dynamic determination of the reconfiguration and measured the power consumption of each P-CSCF while performing calling, call session migration, and turning nodes on/off. First, UEs made calls through four P-CSCFs randomly, and then increased the number of calls to exceed the upper limit of the P-CSCF. We observed the behavior of the reconfiguration caused by the constraint condition. Secondly, the UEs reduced the number of sessions enough to turn off one node and observed the behavior of the reconfiguration. Then, the UEs increased the number of sessions from the reduced number enough to turn on an additional node, and we observed the behavior in the same way. Finally, the UEs cycled through a decrease and increase, and we evaluated the reduction in power consumption compared with the existing static IMS configuration, which was simply traditional operation where no reconfiguration was executed and the four nodes were kept running.

TABLE II. PARAMETERS OF P-CSCFS

params	values	params	values
B_n	48.2 W	$C_{s,t}$	3.10×10^{-3} J*
τ	0.137 W	$T_{s,t}$	1.86×10^{-4} s
ε	1.06×10^{-3} W	h_n	636 J*
K_n	18,000 sessions	W	2 s
U	0.8	R	100 s

* J: W·s (watt second)

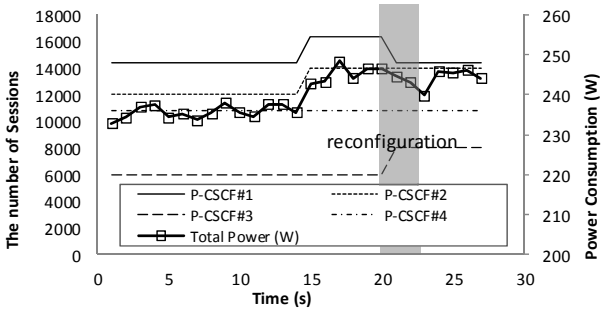


Figure 5. Load Balancing Session

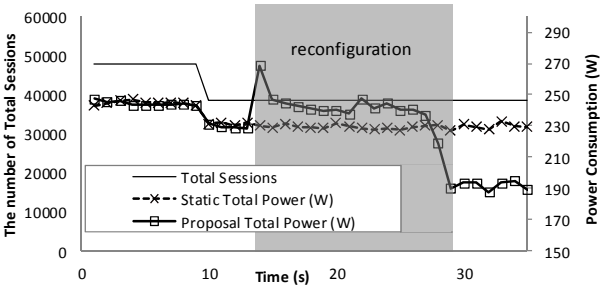


Figure 6. Turning off a node

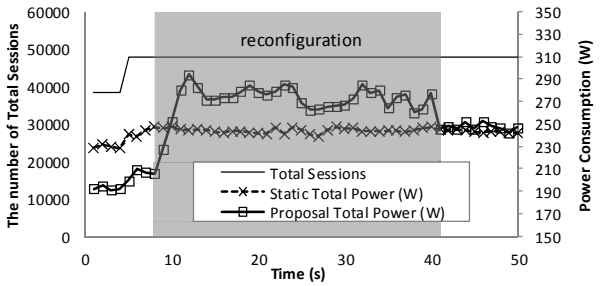


Figure 7. Turning on a node

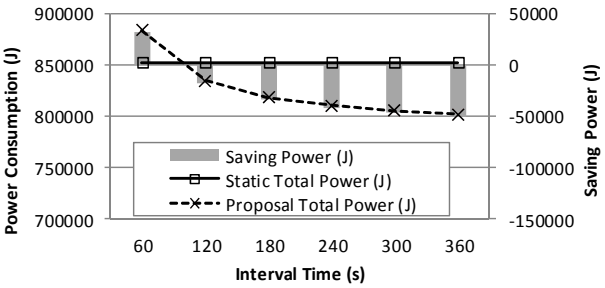


Figure 8. Comparison during reconfiguration cycle

C. Results and Discussion

Fig. 5 shows the load balancing behavior of the reconfiguration when the number of sessions exceeds the upper limit of the P-CSCF, which is 14,400 sessions (18,000 $\times$ 0.8). The reconfiguration began to be executed with periodical timing of calculating criterion in (13). The excess sessions on P-CSCF#1 were migrated to P-CSCF#3. Fig. 6 shows the behavior of the reconfiguration when the number of sessions was reduced. "Static Total Power" in the figure is the power consumption in the static IMS configuration, as in traditional operation. In our proposal, all sessions of a P-CSCF on a node were migrated to other P-CSCFs and the node was turned off. On a temporary basis, its power consumption increased, but it became smaller than the static power consumption after the reconfiguration. Fig. 7 shows the behavior of the reconfiguration when the number of sessions increased. In the beginning, the power consumption in our proposal was small because the number of running nodes was small. After that, it increased temporarily to equal the static power consumption for turning on a node and processing more sessions. Fig. 8 shows the power saving with our proposal compared with the static power consumption while running for one hour. During the period, we repeated the behaviors of Fig. 6 and Fig. 7. The interval time is the time for one cycle of reducing and increasing sessions (R = (interval time)/2). When the cycle of reconfiguration was long enough, although that depends on the situation, the power consumption was reduced by 5.9%. The amount of reduction becomes larger depending on the situation such as the changing number of nodes and sessions.

Thus we verified the value of dynamic determination and reconfiguration to reduce power consumption based on our proposed model and formula for linear programming in a real environment. To use linear programming to obtain an optimized solution, we approximated the non-linear values as accurately as possible with linear values calculated from their averages. This does not have a large impact on our proposal, and could be made more precise by formulating the non-linear values in a more rigorous manner. Also we can apply our proposal to larger configurations because the time to solve the linear problem was about 5 seconds, and larger problems can be solved within practical time constraints.

V. CONCLUSION

We have focused on the dynamic determination of IMS reconfiguration in a flexible IMS environment that can change its configuration by live session migration and turning nodes on/off. In order to realize this, we proposed a model of the reconfiguration and formulation to be solved by linear programming with the goal of saving power. Furthermore, we implemented a dynamic reconfigurable IMS that used our proposal, and verified that the reconfiguration reduced power consumption.

ACKNOWLEDGMENT

This work has been supported by the Japanese Ministry of Internal Affairs and Communications funded project.

REFERENCES

[1] 3GPP TS 24.229 V9.13.0, "IP multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3," 2012.

[2] Global System for Mobile communications Association: www.gsma.com

[3] GSMA, "Rich Communication Suite 5.1 Advanced Communications Services and Clients specification," 2012.

[4] M. Ito, S. Komorita, Y. Kitatsuji, and H. Yokota, "OpenFlow-based Routing Mechanism for Call Session State Migration in the IMS," Proc. 7th WSEAS International Conference on CEA, 2013.

[5] ILOG CPLEX 11.0 User's Manual, 2007.

[6] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, et al., "SIP: Session Initiation Protocol, Internet Engineering Task Force," RFC 3261, 2002.

[7] 3GPP2 X.S0011-001-C v3.0, "cdma2000 Wireless IP Network Standard: Introduction," 2006.

[8] 3GPP TS36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2," 2009.

[9] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, et al., "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Computer Communication Review*, vol.38, No.2, 2008, pp. 69-74.

[10] H. Jiang, A. Iyengar, E. Nahum, W. Segmuller, A. Tantawi, and C. Wright, "Load balancing for SIP server clusters," IEEE INFOCOM, 2009, pp. 2286-2294.

[11] V. Hilt and I. Widjaja, "Controlling overload in networks of SIP servers," IEEE International Conference on Network Protocols (ICNP), 2008, pp. 83-93.

[12] H. Fathi, S. Chakraborty, and R. Prasad, "Optimization of SIP session setup delay for VoIP in 3G wireless networks," IEEE Transactions on Mobile Computing, Vol. 5, Number. 9, 2006, pp.1121-1132.

[13] N. Rajagopal and M. Devetsikiotis, "Modeling and Optimization for the Design of IMS Networks," IEEE ComSoc Proceedings of the 39th annual Symposium on Simulation, 2006, pp. 34-41.

[14] M. Fakhfakh, O. Cherkaoui, I. Bedhiaf, and M. Frikha, "High availability in IMS virtualized network," First International Conference on Communications and Networking, 2009, pp. 1-6.

[15] V. Petrucci, O. Loques, and D. Mossé, "Dynamic optimization of power and performance for virtualized server clusters," Proceedings of the 2010 ACM Symposium on Applied Computing, 2010, pp. 263-264.

[16] The Open Source IMS Core Project, Available at: http://www.openimscore.org/

[17] IMS Bench SIPp, Available at: http://sipp.sourceforge.net/ims_bench/

 ISBN: 978-1-61208-279-0

HARP: A Split Brain Free Protocol for High Availability Implemented in FPGA

Rômerson Deiny Oliveira, Daniel Gomes Mesquita, Pedro Frosi Rosa
Federal University of Uberlândia
Faculty of Computing
Uberlândia, Brazil
romerson@mestrado.ufu.br, {mesquita, frosi}@facom.ufu.br

***Abstract*— High availability is a key requirement for today and future networks. In despite of large investments to achieve high availability, network providers cannot guarantee 100% of availability. The existing protocols have two harmful situations named 'No Brain' and 'Split Brain' conditions, which are algorithmic problems that attack the network availability. This paper aims to show the developing and implementation of a new high availability protocol and how it fixes its predecessors, regarding these conditions.**

Keywords-High Availability; Protocol Design; HARP; VRRP; Split brain.

I. INTRODUCTION

Internet has become one of the most important tools to personal and business transactions in the last years. According to International Telecommunications Union, the total of internet users has increased 389% between 2001 and 2011 [1]. In absolute numbers, this means a rising from 495 million to 2,421 billion of connected users.

Network downtime can bring large financial losses to companies. A published study says that the downtime costs North American businesses collectively $26.5 billion in revenue each year [2]. In order to keep the Internet available as long as possible, universities and manufacturers get started research about high availability.

The HA (High Availability) mechanisms are characterized by using solutions based on hardware redundancy, intelligent software and protocols to identify system failures [3]. This mechanisms work by physical elements seen by the local clients as a single one, named virtual element. The physical elements operate under the master/slave philosophy, such that they have always one master, neither more nor less than one and the others stay available as slaves [4]. The protocol must be able to advertise that there is a master within the group and elect a new one in case of failure on the current master.

The VRRP (Virtual Router Redundancy Protocol) [4] is the *de facto* standard protocol to HA equipments such as routers and switches, mainly in telecommunications networks. Despite all efforts to keep the network available, downtime intervals were noticed onto networks with VRRP in operation [5] and two harmful situations have been identified as responsible for these issues. The first situation is named 'split brain', while the second one is the 'no brain' condition (Subsection II.A).

Hashimoto et al. [5] presented a proposal extending the VRRP protocol to solve the addressed problems. The proposal was modeled in Petri net [6] and due its high level of abstraction, it was necessary refine it and define the five protocol elements [7], intending its implementation.

From this scenario, this paper aims to present a new HA protocol, named HARP (High Availability Router Protocol), free of such phenomena, by defining their elements and show how it fixes its predecessors. The protocol analysis is focused on their proof of concept. We bring data about hardware operation, but a deeper analysis about HARP performance will be presented in future works. The HARP elements, namely assumptions about environmental, services, vocabulary, formatting and procedure rules are presented in Section IV.

The HARP is part of a bigger project that researches Future Internet conducted by MEHAR research group and deals with the HA aspects of the EDOBRA project, that in turn, intends to expand the physical coverage of the experimental installation OFELIA in Brazil [8]. The HARP's first version is addressed here, developed to the current Internet architecture.

Future versions as IPv6 and clean slate compliant are not covered; however, there are researches in progress about these versions and they can be prototyped, once the reconfigurable platform chosen to prototyping. Furthermore, statefull protocols are out of the scope of this work, which is related with stateless protocols like IP protocol.

The remainder of this paper is organized as follows: Section II shows a summarized HA background and related work. In Section III, the method of development to HARP and the set of steps to its validation are presented. Section IV details the protocol analysis, and lastly, Section V shows the conclusion and potential future works.

II. RELATED WORK AND BACKGROUND

This section covers the basics of HA mechanisms and situations which affect the protocols. Following, there is a review of related work, showing the relation with this one.

A. High Availability Overview

HA refers to the network ability to remain available close to 100% of the time, preventing loss of service by reducing or managing failures and minimizing unplanned downtime for the system. HA is obtained by using a virtual address shared by two or more NEs (Network Elements). This address is defined to be the default network gateway for

 ISBN: 978-1-61208-279-0

internal hosts. The virtual router is an abstraction, which consists of one or more routers running a HA protocol. If a failure occurs in the main NE of the group (Master), another HA group's component takes its function, using a virtual IP address corresponding to a virtual MAC (Media Access Control). Because of this, the handover of network elements becomes transparent to local clients, since communication remains on [9].

HA is a subset of fault tolerance since the last one ranges from hardware redundancy up to communications management with a protocol. From the protocol aspects, there are two main causes that lead network to outage: (1) 'no brain' condition in which the master becomes out of service and the infrastructure has no other nodes (slaves) able to assume the master role, i.e., indeed, there is no routing of packets for a while, and (2) 'split brain', which is the situation in when two or more infrastructure nodes assume the master role, i.e., there are two or more routers sending the same packet forward by resulting in replicated requests, which will lead the transport protocol to unpredictable situation) and to the inaccessibility [5].

These conditions can be caused by interface failure (resulting losses or errors in messages) or attack by third parties. This work applies to the first issue.

B. Related Work

The literature provides several publications that can be related to this. The one hand, there are FPGA implementations applied to computer networks and fault tolerance at hardware level. On the other hand, there are investigations concerning protocols at higher levels of abstraction.

Jiang and Prasanna [10] explore the abundant parallelism of FPGA [11] to handle the new generation of packets classification and propose improvements in this regard, in order to make the packets classification and forwarding more scalable free from losses in transmission speeds. This feature is included in the HARP architecture and will be essential when different HARP versions are prototyped in the same core.

Casado *et al.* [12] show that use specialized hardware for packet forwarding is an efficient technique. Also in [12], it is inferred the idea of using more flexible network processors as a way to avoid redoing chips due to changes in protocols or add new features to this hardware, given that cost is a limiting factor.

Straka and Kotasek [13] present a methodology for building fault-tolerant systems based on FPGA. The architectures are based both on technique duplex system as in triple modular redundancy to improve fault detection. For this purpose, the use of testers on-line is shown. It is also shown how the parameters of availability (e.g., mean time between failure and recovery rate) may be affected by operating environment in which the fault tolerant system is implemented. The work is focused on hardware replication techniques and brings methods for building redundant hardware elements, whereas the work shown in this paper handles the communication between the elements.

Lopes Filho [14] examines the idea in that transport layer could be the main problem of high availability protocols, due to using connectionless protocols, resulting in a proposal of a transport layer based on the SCTP protocol. However, the author concluded that the problem lay not in the transport layer and the suspicion came to link layer, due to broadcast messages with false positive for the upper layers.

Hashimoto [15] pointed the errors not detected by the link layer control algorithms as cause for the split brain problem and concluded that would be necessary develop a new HA protocol, or even redesign an existing one. It was shown that the VRRP automaton is incomplete, once it does not take in account errors not detectable by the link layer. The author defends that the problem encountered in VRRP also applies to CARP and HSRP protocols. The author defends that the problem encountered in VRRP also applies to CARP and HSRP protocols. The work presents a Petri net specification, which foresees the loss of advertising messages depending of circumstances in the Link layer.

Pereira Junior [16] discusses the conditions for the concurrent no brain and split brain situations and defines a service specification that composes the HA protocol design. The thesis presents assumptions for an environment where a HA service should operate and how HA protocols could fix arising challenges from there.

This paper stands as a continuation of the research developed in [14][15][16]. These three, in turn, describe a sequence of hypotheses and conclusions about the problems that attack HA protocols and are summarized in [5]. The authors concluded that the problems lie in the no brain and split brain conditions, realized in the VRRP protocol.

III. METHOD

The construction process of the High Availability Router Protocol may be sorted in three stages. At first, we get started by analyzing a proposal of VRRP extension, modeled in high level of abstraction. Specification and development of the HARP elements came following. Lastly, we propose an evaluation system for HA protocols.

A. Characterising the problem

Unavailability situations had occurred onto a network with VRRP in operation [14]. In certain conditions, the finite state machine (FSM) generated by this protocol presents the split brain situation. In other conditions, it presents the brainless one [5].

Two protocols have been largely accepted regarding this matter: VRRP [4] and CARP (Common Address Redundancy Protocol) [17]. The first one has been developed as a proprietary protocol and the second as a protocol developed by the free software community. There also the HSRP (Hot Standby Router Protocol) [18] to provide HA. The finite state machine generated by these protocols presents the same problem in their FSM [5][15].

After finding out the downtime causes and point them out as algorithmic problems that attack the protocol, in [5] was presented an abstract specification, modeled on Petri nets and represented by an automaton to circumvent the cited problem. The automaton in [5] has not been implemented

 ISBN: 978-1-61208-279-0

and it is a description in the same Petri net specification philosophy: in view of the whole HA group's behavior (virtual element).

However, in [5], there is no individual behavior specification to each one physical element, which forms the virtual element. In this way, it would be needed refine the proposal in order to implement it and validate the VRRP extension and because of this, also specify the automaton and the four remaining elements to each protocol instance [7].

This paper brings the behavior rules for a HARP instance in a physical element, from its FSM viewpoint, which was lacking in [5]: items such as parameters and conditions for the master election process, input and output conditions considered in transitions between states for an instance, service specification, and foresee losses of primitives were still necessary for an implementation. This refinement process yielded the HARP.

B. Bulding the High Availability Router Protocol

In this phase, the HARP specification and the creation of its elements were done, as well as its hardware implementation. The five protocol elements are presented with details in Section IV, while the implementation aspects are treated here.

Hardware implementation allows detect some details and correct failures which can remain unnoticed when we are designing protocols at high levels of abstraction. Nevertheless, the production costs to have a specific hardware is too large. In this context, use a reconfigurable hardware could improve the results and reduce costs, so that FPGA is the most indicated for this purpose.

By using FPGA, it was possible to make corrections after each HARP prototyping. This process allows find hardware errors, come back to review the specification and fix them. This iterative sequence makes possible to specify the HARP and prototype a specific hardware till we get a final version. The FPGA usage to implement network algorithms has been widely used in enterprises and universities due to its adaptability and flexibility, reducing production costs as well as time-to-market when compared to an ASIC.

A platform was built in order have a proof of concept of the HARP. For that, a scheme was made with three prototyping boards simulating network elements, connected each other over a star topology. Each NE has an FPGA Cyclone 2 [19] connected as shown in Figure 1.

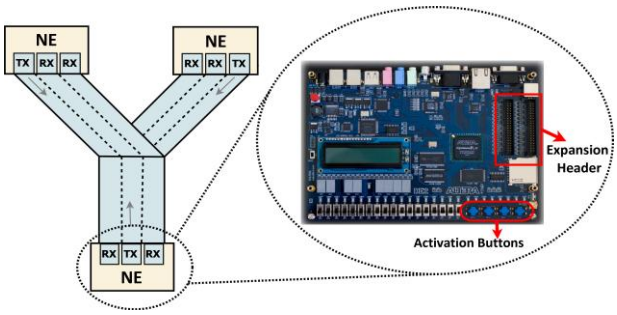


Figure 1. Proof of concept plataform

Each chip relays data to any other one through dedicated connections on a flat cable of 40 paths connected to the expansion header into the board (36 paths for data and 4 paths for supply). Service requests are done by pressing the activation buttons on the board. The channel was divided in three paths with 12 wires each one.

The activation buttons trigger the services shown in Table 1. Each service was tested and after a set of hardware reconfigurations, they worked correctly. Since HARP is based on timeouts (Subsection IV.D), control delays into the states and attempt to all possibility of losses of primitives were the main benefit due to FPGA reconfigurability.

HARP was tested foreseeing the loss of each recognized primitive in its vocabulary. Thus, it possible to guarantee that HARP can deal with all situations of loss of primitives. HARP was tested foreseeing the loss of each recognized primitives in its vocabulary.

Thus, it possible to guarantee that HARP can deal with all situations of lost primitives. This analysis was done concurrently to the implementation. Some examples of loss of messages are given in Subsection IV.D.

As mentioned previously, Figure 1 represents the proof of concept to the protocol elements and it was done successfully, but it did not address the system time to recovery neither bring data about how long each primitive takes to being processed into the HARP core. To treat this, we propose a validation system in the next section, capable to test any HA protocol.

C. Validation System

Figure 2 shows the proposed validation system to test HA protocols. It has a virtual element composed by three (no limited to) network elements (FPGA) [19] connected each other through a shared channel. This channel is connected to PCs responsible for generate information flow.

The system checks HA protocols from the viewpoint of interface and channel failures, since this is the main prompter of no brain and split brain conditions. Another possible cause is the network attack by third parties, but it is out of this work scope.

Each one of NEs in Figure 2 is configured as a System on Chip (SoC) centered on an embedded processor. The HA protocol may be one SoC component, when implementet directly in hardware, or may be a portion of the application running over the SoC.

The SoC should have at least a processor, clock generator and memory core. To test the protocol prototyped in

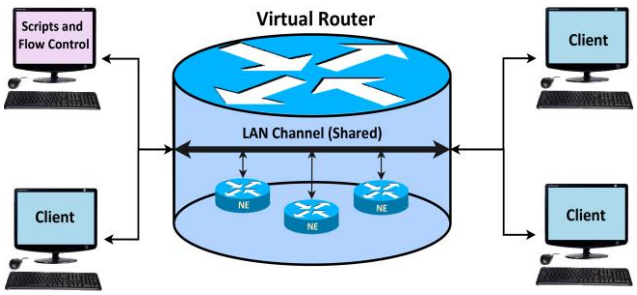


Figure 2. Validation System model proposed

 ISBN: 978-1-61208-279-0

hardware, it has to be included as a SoC module and a hardware abstraction layer must be provided to enable the communication between the protocol and the system interconnection bus. To test the protocol in software, it must be added as a portion of the application source code. The application has to be formed by Ethernet, ARP and IP cores. PCs generate flow information by using ICMP messages.

To test HARP, it was included as a hardware module in a SoC. In normal operation, the master Ethernet channel was intentionally crashed. After that, the Wireshark software [21] shown the message exchanging in the PC Ethernet ports, as well as the embedded application showed the packets in the SoC Ethernet port. Hence, we can control the whole election process and get the advertisement messages from the new master.

Prototyped in an Altera Cyclone II EP2C35F672C6 FPGA, the HARP had used a small portion of available resources. The occupied area reached the mark of 9.2% of the 33216 logical elements (LE). For testing purposes, it was synthesized to an Altera Stratix IV EP1S80F1508I7 FPGA. In this second case, the total used area reached only 1% of the 182400 ALUTs available in the FPGA [20].

The actual HARP performance evaluation is based on the circuitry [20]. The maximum frequency reached by the HARP hardware module is 92.68 MHz. From this point, it was established a ratio between the HARP frequency and the time required to transmit a bit in different channel speeds. This ratio is described in [20] and, summarily, it takes in account the size (bits) of a HARP message and the minimal time required by the channel to transmit this amount of bits.

Actual traffic has been considered in a production scenario to verify the error rate in a local high availability infrastructure. Regarding HARP, only a proof of concept has been constructed, as we don't have how to reproduce real telecommunications network traffic. An ongoing research is preparing an environment to couple HARP with other TCP hardware modules and submit it to a real traffic in a partner telecommunication company.

IV. HARP ELEMENTS

The five protocol elements are presented in this section: assumptions about environmental, services, vocabulary, formatting and procedure rules. To save space, services and vocabulary will be shown together.

A. Assumptions about Environmental

HARP, as a HA protocol, must to operate in each single element into a group of redundant ones. This group forms the virtual element, which is seen by the local clients as a single point of packets forwarding. Particularly, HARP is projected to operate into network layer of the Internet architecture, coupled with Internet Protocol.

B. Services and Vocabulary

In this section, there is a presentation of all services provided by HARP. The primitives are also introduced according to the Request, Confirm, Indication and Response taxonomy. Table 1 summarizes the services set and specifies what messages are exchanging during its execution. Subsection IV.D resumes this topic and provides more details about services execution and the usage of messages.

There are also two messages not cited in the Table 1, given that they are not part of specific services, they are:

- Active Slave Request (ACTS_REQ): sent by the master, in broadcast, intending update its active address table.
- Active Slave Response (ACTS_RESP): sent by a slave, to the master, confirming its activity

C. Message Format

In order to attend the services requirements and also by operate on network layer of Internet architecture, HARP messages have ten fields defined in its first version, as shown in Figure 3.

TABLE 1. HARP SERVICES AND VOCABULARY

Service	Messages	Description
Keep Alive (KA)	*KA Request (KA_REQ)*	Unconfirmed service. It acts as heartbeat and is used by the master to advertise its availability. By monitoring these heartbeats, the slaves determine when a master instance has stopped.
Given Master (GM)	*GM Request (GM_REQ)*	Confirmed service. Used by the master to indicate to the others nodes that it is willing to transfer its role. The target slave address must be included in the active address table.
	GM Response (GM_RESP)	
	GM Failed Request (GMFAIL_REQ)	
	GM Ready Request (GMRDY_REQ)	
Inform Node (INF)	*INF Request (INF_REQ)*	Confirmed service. A service used to indicate that a new node is becoming a member of the HA group.
	INF Response (INF_RESP)	
Remove Node (REM)	*REM Request (REM_REQ)*	Confirmed service. A service used by a slave node, to indicate to all other nodes of the group, that it is leaving the HA group.
	REM Response (REM_RESP)	
Check Brain (CB)	*CB Request (CB_REQ)*	Confirmed service. A service used by a slave node to certify itself that there is no master node in the group HA and avoid the split brain situation during the master election process.
	CB Response Positive (CB_RESP(+))	
	CB Response Negative (CB_RESP(-))	

The messages have 128 bits of header in this version compliant with IPv4. HARP messages are mainly control ones and then have the header bigger than the data field. By the way, DATA field is not used yet, but can be used in the next versions. Each field is explained following.

- DEST_ADDR: target IP address. It may be unicast or broadcast;
- SRC_ADDR: source IP address. It says what NE sent that message;

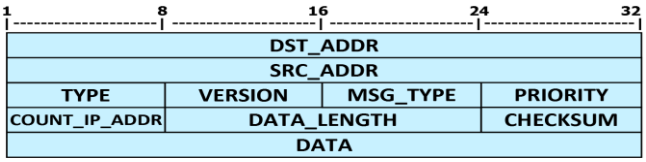


Figure 3. Message Formatting

- TYPE: protocol type, i.e., reserved number to the HARP identification;
- VERSION: HARP version, in this case it is one;
- MSG_TYPE: this field indicates what type of HARP message is being transmitted;
- PRIORITY: NE priority, set up by the network manager;
- COUNT_IP_ADD: in a KA_REQ, indicates the total of slaves in the HA group, can also be used to tell how many addresses are into the data field;
- DATA_LENGTH: it means how many bits are carried in the DATA field;
- CHECKSUM: verification code to detect errors;
- DATA: field designated to carry message parameters when necessary.

The 0x0000 and 0xFFFF addresses are reserved. The first one means that nothing is in the address field, whereas the last one mains broadcast. Next HARP versions can have not identical messages to these ones, in order to serve different network architectures.

D. Procedure Rules

Procedure rules explain the messages exchange and the protocol behavior. Intending to express the HARP behavior unambiguously, represents the HARP FSM. Each service provided by HARP has a couple of automata representing sender and receiver. The FSM in the is the union of each single automaton (according to services) considering sender and receiver instances. The final automaton has eight states. This section will clarify the transitions between states, linking them to each service.

Some states have timeout intervals as input in their transitions. For a while, they are just introduced here, its usage is cited over the text. There are five kinds of timeouts based on a time interval *t*, whose values are: $TO_1 = t$; $TO_2 = t$; TO_3 = (2+Priority)**t*; $TO_4 = t$ and $TO_5 = t$ (sender) or $2t$ (receiver).

In the case of loss of messages, the HA group can be driven to no brain or split brain conditions. A set of simulations are conducted along this section in order to show how HARP fixes this harmful situations, especially the split brain, which can be triggered by distinct actions.

1) Keep Alive (KA): The node starts at *Idle* state. Based on priority (zero for master and higher to slaves), it will from *Idle* state to *Master* state or to *Slave* state. If a node assumes *Master* state, it sends periodically the KA_REQ by informing its state. Master uses the COUNT_IP_ADDR field to inform how many slaves are in the group.

2) Inform Node (INF): When a NE wants to come into the network, it sends an INF_REQ and waits for a INF_CONF from the master. The master will include its address into the active address table. A master only needs to send INF_REQ when it becomes a slave, but does not need wait for confirmation, since the connection is already established.

3) Remove Node (REM): when a NE sends the REM_REQ it means that it should be removed of the active address table. The requesting node will leave the network only after receiving the service confirmation REM_CONF sent by the master node.

4) Given Master (GM): The master NE sends GM_REQ and transits to *Wait For Given Master Confirm* state (WF_GM CONFIRM). If it receives a GM_CONF, the NE goes to slave state and sends a GMRDY_REQ informing the process conclusion. However, if TO_1 happens, it sends a GMFAIL_REQ informing the receiving error and the transition is to *Master* state again.

At the receiver side, when the slave NE receives a GM_IND, it sends a GM_RESP and goes to *Given Master Accepting* state (GM_ACCEPTING), waiting an interval to ensure there is not NE sending Keep Alive messages. If a GMRDY_IND arrives, the transition is to *Master* state, but if there is a TO_2, a KA_IND or a GMFAIL_IND, the receiver returns to *Slave* state, avoiding the split brain condition.

Some situations are illustrated within a GM context in order to show the HARP ability to recover itself:

- If GM_REQ is lost, the receiver does not receive GM_IND and neither replies with GM_RESP, then the master comes back to Master state at TO_1 and sends a GMFAIL_REQ. Assuming that TO_3 is the interval to realize the master failure and start an election process, TO_1 must be lower than TO_3, ensuring that the will not occur no brain situation;
- In the case when GM_RESP is lost, GMRDY_REQ will not be sent by the sender, so the receiver will not go to *Master* state. The receiver, currently at GM_ACCEPTING state returns to Slave state when TO_2 occur;
- If GMFAIL_REQ is lost, there is no worry, since TO_2 or an incoming KA_IND cover this situation;
- When GMRDY_REQ is lost, the sender now is already a slave and the receiver will also return to slave. This is the worst case within a GM context, but the Check Brain process will be consequently started and fixes this situation.

5) Check Brain (CB): If the slave detects a master absence, a process to elect a new master is started. This is a crucial point of the protocol. The CB_Flag is used to inhibit that the CB process being started by more than one slave at the same time.

If there is a TO_3 without KA_IND, the slave NE sends a CB_REQ to all others slaves by ensuring there is no master and goes to *Wait For Check Brain Confirm* state (WF_CB_CONFIRM). By arriving a CB_CONF(+) or KA_IND or happening a TO_4, it returns to *Slave* state, knowing there is a master onto HA group. But, by arriving a CB_CONF(-), it transits to *Master Election* state to wait for more negatives confirmations, ensure the master failure and changes to *Master* state.

On the receiver's side, when a CB_IND is received, the slave NE goes to *Search Master* state, which verifies the KA_IND receiving. If this reception has been successfully, a CB_RESP(+) is sent and a transition returning to *Slave* state happens, otherwise an CB_RESP(-) is sent and the this NE

 ISBN: 978-1-61208-279-0

also going to *Master Election* state (which forces it to wait for $TO_5 - t$ for sender and $2t$ for receiver, while the sender concludes the CB process).

TO_3 interval is based on node priority in order to be different in each one of them, so minimizing the probability of having more than one slave beginning the election process at the same time. Even so, if two slaves start the process concurrently, the other nodes will respond the first message that arrives to them. O interval $2t$ with no KA_IND is already enough for a slave reply a CB_REQ with CB_RESP(-), to ensure the slave with lowest priority can also respond.

If a master becomes unavailable, a new master is elected and then the previous one returns to master role, the split brain may occur. To fix it, if a master receives a KA_IND, it goes to *Slave* state, if the two masters go to *Slave* state, the Check Brain process is started.

The nodes in a HA group belong to a multicast group. Master sends KA_REQ messages only in this domain. Hence, switches can control the message flow to do not reach end systems. The same configuration happens about the CB_REQ message. It can be transmitted only in a multicast group. The rest of the HARP messages are exchanged in unicast mode and, because of this, there no discussion about to go out on all end systems.

To make HARP compatible with IPv6, a new message format must be done. The vocabulary will be hold as the actual, but the fields have to be revised. By keeping the vocabulary, there is no need to increase the message size, but merely reorganize the fields. Philosophically, it will not be performed any change in the protocol, regarding IPv6 compatibility. On the other hand, the prototyped architecture needs to be reconfigured, intending to keep the exploration of spatial parallelism and the processing frequency.

In this work, we were concerned in demonstrate that there was a misconception in current high availability protocols like VRRP and CARP. HARP is a free split brain protocol, but security issues must be opportunely addressed. For example, we could introduce some confidentiality and authentication to avoid the issue mentioned above. A foreseen assumption is respect introducing the tamper-resistant and tamper-evident environment to the whole HA core circuit environment [22].

E. Final Automaton

Thus, by combining the service automatons, it is possible generating a global automaton which expresses whole protocol instance's behavior. Figure 4 shows all events and actions arising from this. Description of all states in the FSM was seen at Subsection D, as well as all recognized events considered to transitions between states. This FSM brings the HARP behavior to any provided service.

The states S5, S6 and S7 (Figure 4) are the main states to be visited intending keep the HARP free of the Split Brain condition. In case of receiving HARP messages with errors, the election process are invoked and the conversation between the HARP instances must confirm that more than one equipment do not see the master. This avoids that a slave becomes master when it owns the punctual failure.

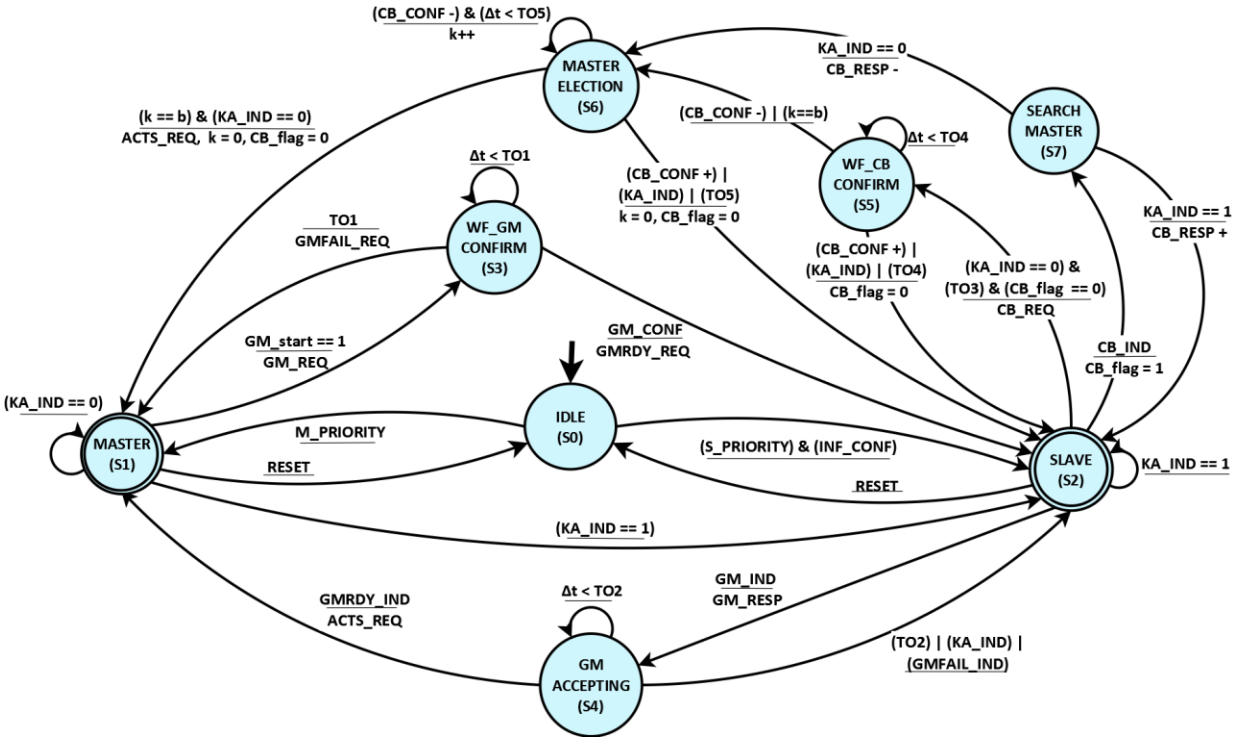


Figure 4. HARP Finite State Machine

The automaton is well formed and keeps the good properties: it is k-limited, guaranteeing the FMS being finite; It does not have states that do not present state thereafter, i.e., free from deadlock; is free of live-lock, i.e., it ensures that there is not a sequence of states that has no successor state; and there is always a sequence of transitions/states that lead to the initial state. So, the modeling was done successfully, guaranteed by complete description of its elements and demonstration of good properties.

V. CONCLUSION AND FUTURE WORKS

The main contribution of this paper is presenting the High Availability Router Protocol and how it fixes the algorithmic problems that attack the existing high availability protocols. It is important to remark that VRRP, a leading protocol of the telecommunications industry, has been the benchmark.

To get HARP, we started from a proposal at high level of abstraction, modeled with Petri net. Thus, we made a complete specification at a lowest level of abstraction to the HARP, without escaping crucial issues to implementation.

The text shows, especially by analyzing the HARP automaton, that HARP maintains the FSM good properties: it is k-limited, free from livelock and deadlock and it is re-initializable from any state.

The two main problems that attack high availability are the no brain and split brain conditions. These problems are fixed by the HARP, considering that the new FSM has enough states and transitions to cover these possibilities. Furthermore, QoS parameters could be considered in order to elect the actual router to support a specific traffic. Apart HARP has a complete FSM, it introduces a service that allows transferring the master role spontaneously.

FPGA implementation allowed executing updates without costs, due to reconfigurability. Thus, it was possible change the hardware even after each prototyping. It should be remembered that performance and timing parameters were not measured till the current version. This phase aimed demonstrate HARP correct functioning. Nevertheless, HARP can operate in Gigabit network with its current 92.68 MHz frequency, reached with a low cost FPGA [20], keeping the protocol convergence time pretty lower than one second.

The time to realize a master failure is actually 3 seconds in the existing HA protocols. HARP brings an adaptable time, depending on the time to transmit a message in the system. HARP default interval to realize a failure is 100 ms. It can be higher if in the system, one message takes more than 100 ms to be transmitted between HARP instances.

One issue can be addressed as disadvantages to use HARP. It is still missing the considerations about state transfer to warranty the failover. In addition, there is no specification about load balancing.

This paper presents results about a new network protocol. Such results were built after an iterative process of tests and prototyping in reconfigurable hardware. From now on, we can foresee the development of a specific hardware based on HARP to deal with high availability network. As potential future works, we suggest (1) implement the validation system, considering TCP/IP protocols in the network and link layers and (2) prototype HARP architecture to be compliant with IPv6 and clean slate approaches.

REFERENCES

[1] International Telecommunications Union Internet users. <http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx> 05.06.2013

[2] C. A. Technologies. <http://www.ca.com/us/news/Press-Releases/na/2010/North-American-Businesses-Lose-26-5-Billion-Annually.aspx> 05.06.2013

[3] R. Cameron, B. Woodberg, M. K. Madwachar, M. Swarm, N. R. Wyler, M. Alber, and R. Bonnell, Configuring Juniper Networks NetScreen & SSG Firewalls. Rockland: Syngress Publishing, 2007, pp. 587-589

[4] R. Hinden, "Virtual Router Redundancy Protocol". RFC 3768, 2004.

[5] G. T. Hashimoto, E. L. Filho, J. E. Pereira Junior, and P. F. Rosa, "High avaliability: A long-term feature in network elements". Fifth International Conference on Systems and Networks Communications (ICSNC), IEEE, 2010, pp. 201 –206.

[6] C.A. Petri, Kommunikation mit Automaten, Ph.D. Thesis. University of Bonn, Germany, 1962.

[7] G. J. Holzmann, Design and validation of computer protocols, New Jersey: Prentice-Hall, 1991, p. 21.

[8] MEHAR (2012). Extending and deploying ofelia in brazil. <http://www.mehar.facom.ufu.br/projects/ofelia-edobra.dot> 05.06.2013

[9] J. Sonderegger, O. Blomberg, K. Milne S. and Palislamovic, Junos High Avalability: Best Practices for High Network Uptime. USA: O'Reilly Media, 1st edition, 2009

[10] W. Jiang and V. Prasanna, "Scalable packet classification on fpga". IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol 20, no 9, Sept. 2012, pp. 1668 –1680.

[11] S. Brown. and J. Rose, "Fpga and cpld architectures: a tutorial", Design Test of Computers, IEEE, vol.13, no.2, 1996, pp.42,57.

[12] M. Casado, T. Koponen, D. Moon, and S. Shenker, "Rethinking packet forwarding hardware", In ACM Workshops on Hot Topics in Networks, volume VII, 2009, pp. 1–6

[13] M. Straka and Z. Kotasek, "High availability fault tolerant architectures implemented into fpgas". In Digital System Design, Architectures, Methods and Tools, DSD '09. 12th Euromicro Conference, 2009, pp. 108 –115.

[14] E. Lopes Filho, "Arquitetura de alta disponibilidade para firewall e ips baseada em sctp". Master Thesis, Federal University of Uberlândia, Brazil, 2008.

[15] G.T. Hashimoto, "Uma proposta de extensão para um protocolo para arquiteturas de alta disponibilidade" Master Thesis, Federal University of Uberlândia, Brazil, 2009.

[16] J. E. Pereira Júnior, "Especificaçãoo de serviço e suposições sobre o ambiente para um protocolo de alta disponibilidade". Master Thesis, Federal University of Uberlândia, Brazil, 2010

[17] OpenBSD. Pf: Firewall redundancy with carp and pfsync. <http://www.openbsd.org/faq/pf/carp.html> 05.06.2013

[18] T. Li, B. Cole, P. Morton and D. Li,. Cisco Hot Standby Router Protocol. RFC 2281, 1998.

[19] Altera. Development Education Board DE2 <http://www.altera.com/education/univ/materials/boards/de2/unv-de2-board.html> 05.06.2012.

[20] D. G. Mesquita, R. D. Oliveira, and P. F. Rosa, "The High Availability Router Protocol in FPGA," unpublished.

[21] Wireshark. <www.wireshark.org> 05.06.2013

[22] G. E. Suh, D. Clarke, B. Gassend, M. V. Dijk, and Srinivas Devadas. "AEGIS: architecture for tamper-evident and tamper-resistant processing", In Proceedings of the 17th annual international conference on Supercomputing (ICS '03), ACM, 2003, pp. 160-171.

 ISBN: 978-1-61208-279-0

Employing the CEP Paradigm for Network Analysis and Surveillance

Ruediger Gad, Martin Kappes
University of Applied Sciences Frankfurt am Main
Frankfurt am Main, Germany
e-mail: {rgad, kappes}@fb2.fh-frankfurt.de

Juan Boubeta-Puig, Inmaculada Medina-Bulo
Universidad de Cádiz
Cádiz, Spain
e-mail: {juan.boubeta, inmaculada.medina}@uca.es

***Abstract*—In this paper, we present a network analysis and surveillance system based on the Complex Event Processing (CEP) paradigm. We demonstrate how complex event hierarchies based on single packets can be leveraged for detecting attacks such as, e.g., SYN Flooding, and present experimental performance results indicating that current CEP implementations running on consumer class computers are well capable of analyzing network traffic volumes with such patterns in the Gigabit range, rendering our approach applicable for enterprise network monitoring.**

Keywords—CEP; network analysis; network surveillance.

I. INTRODUCTION

Today, Information Technology (IT) and its underlying computer networks are the foundation of virtually all business infrastructures. Since mission-critical processes depend on the reliable operation of the IT, the continuous operation of the network has to be ensured. Thus, the quality, reliability, performance and serviceability of the underlying computer network are paramount. Consequently, systems for assuring the proper operation of computer networks are needed.

Computer networks resemble, in many ways, a complex nervous system interconnecting individual nodes. Their operation is affected by a broad range of factors such as, but not limited to, e.g., component failures, traffic overload, misconfigurations, or attacks. In order to assure the proper operation of a computer network, situations which negatively affect the network operation must be rectified in a timely manner. Ideally, critical situations are avoided altogether by detecting and reacting to conditions that may eventually become critical beforehand.

Thus, information about relevant situations is needed. In terms of computer networks this means that data about the computer network and especially the occurrences in it is required. Hence, the very first step for assuring the proper operation of a computer network is the collection and analysis of data about the computer network. This data is the very basis for all other subsequent activities like the notification of administrative personnel or the implementation of counter measures.

Gathering and analyzing data in today's complex computer network structures is a challenging task. In order to acquire a comprehensive overview it is insufficient to collect and analyze data at a single point with standalone tools. Instead, a network analysis and surveillance system must be capable of distributed operation, allow for the integration of heterogeneous data sources, and to quickly and flexibly grow or adapt to new situations.

Ideally, the general system structure will be suitable for an extension beyond network analysis and surveillance and support the integration of methods for reactions into the overall system. This way, perspectively, a unified network management and security infrastructure can be implemented that supports the whole network operation, maintenance, and security process.

Consequently, we chose a system approach that is very flexible, not bound to a specific application domain, and allows for a simple extension. Our approach is based on the Complex Event Processing (CEP) [1] paradigm. In this paper, we present the general idea of our approach to leverage CEP in the field of computer networks and analyze its practicability by exemplarily implementing methods for detecting different situations in computer networks and measuring the performance. While CEP had been already used in specialized and isolated areas in the application domain of computer networks, to the best of our knowledge, no overarching, general and unified system had been proposed yet.

The remainder of this paper is organized as follows: at first, we give a summary of different network analysis and surveillance approaches. Then, we give an overview over work related to our research followed by an introduction of the relevant technologies in the fields of network analysis and surveillance and CEP. Afterwards, we introduce our approach and present the prototype that we used to assess the feasibility. Following, we describe the approach we used to assess the feasibility of using CEP in the field of network analysis and surveillance and present and discuss our results. Finally, we conclude this paper with a summary and outlook on future work.

II. RELATED WORK

While other approaches on employing CEP and event-driven architecture (EDA) exist in the field of computer networks, these approaches are typically employed in specialized and isolated application fields only. However, they demonstrate that the CEP paradigm, in general, can be effectively applied in the application domain of computer networks.

Some approaches aim on implementing intrusion detection systems based on CEP [2][3]. These approaches solely focus on detecting intrusions and can be viewed as evidence that CEP is capable of modeling important situations in the area of computer networks. Our approach has a wider scope; we use CEP for general network analysis and surveillance. This includes the detection of intrusions,

 ISBN: 978-1-61208-279-0

but also other use cases like network monitoring or detecting congestion situations, misconfigurations, or faults.

In [4] and [5], a joined infrastructure for detecting attack preparations like so called "stealthy port scans" based on CEP is proposed. This research shows the suitability of modeling computer network incidents by means of event patterns and illustrates the potential of CEP and EDA for creating distributed network surveillance systems. However, [4] and [5] are focused on a single, very specific application while we propose to apply the CEP paradigm in the whole field of network analysis, surveillance and eventually reaction and mitigation.

Another line of research considers CEP-based systems for network management [6]. This work primarily focuses administrative aspects and largely ignores technical problems like data acquisition, data analysis, or triggering reactions. In the long term, we plan to create a unified network analysis and surveillance system that enables the seamless integration of other components like management systems or reaction systems.

In [7], an approach on using CEP for anomaly detection in computer networks is presented. This work shows the potential of using CEP-based systems even for advanced analysis methods like anomaly detection. However, this work is limited on anomaly detection. In our approach, we plan to integrated a multitude of possible analysis methods including classical, pattern-based approaches, as well as machine learning and anomaly detection techniques.

Performance aspects, both in the field of network analysis and surveillance as well as in the field of CEP had also been subject to research. In [8], a cooperative approach on capturing and processing packets at wire-speed in computer networks is proposed. In this work, the load of capturing and processing packets in a network is distributed among multiple capturing components. This approach is not implemented with means of CEP and thus is not as versatile and flexible as our proposed approach.

In [9], a hardware-based CEP engine is presented. With this "in-hardware" CEP implementation a throughput of 20 Gbps could be achieved. This shows the performance potential of CEP systems and that CEP systems can be implemented in-hardware in order to increase the performance. Here, we use general-purpose Java implementations. However, depending on the requirements, special hardware-based solutions may become viable solutions as well.

There also exist well-known monitoring tools like Nagios and Zabbix [10][11]. These tools emphasize on the hardware and services and not on the network itself. Devices and parameters that are usually monitored include performance, like CPU and memory, disk space, temperature, databases, or power systems. Furthermore, typically, such tools display the gathered data in large dashboard-like views. As we will discuss in the presentation of our approach, we believe that it is very important to support the administrative personal by inferring and extracting the important information instead of overwhelming them by displaying all available information. Yet, the information gathered by such monitoring systems is also very valuable. Hence, we plan to incorporate such monitoring solutions in our system as well. Our chosen architecture already supports the flexible and effortless extension of our system and integration of different data sources and provides mechanisms for meaningfully processing heterogeneous data from different data sources.

III. NETWORK ANALYSIS AND SURVEILLANCE

In order to maintain a reliable, robust, and operational computer network, information about the computer network and the occurrences in it must be collected and analyzed in order to identify issues early on. Apart from noticing problems that already affected the operation of a network like a defective hardware component, continuous monitoring of a computer network also enables to discern trends like, e.g., an increase in network traffic or error rates over time, and thus aids in identifying and resolving issues before they become critical. For example, an upcoming congestion situation could be extrapolated by a monitoring system long before the actual effects impair the network functionality as suitable measures can be prepared and implemented early enough. Additionally, continuous monitoring also allows for identifying suspicious activities which may be related to ongoing attacks or attack preparations.

Thus, the very first step in assuring the proper operation of computer networks is to gather information about the network and the occurrences in it.

The act of gathering this data is referred to as network analysis and surveillance or network reconnaissance [12]. While usually the analysis of the data and the deduction of meaningful information is also usually considered part of network analysis, surveillance or network reconnaissance, we primarily consider the data acquisition part here.

Data acquisition can be distinguished into passive and active approaches [13]. Passive data acquisition is based on observing the network only whereas active data acquisition also generates network traffic, e.g., by attempting to contact a device. Here, we only consider passive approaches due to space limitations.

Examples for activities that may be observed in computer networks include packets being transmitted, connections being established or torn down, attackers scanning a network, hosts being added or removed from a network, or services being announced, requested, or used. While there is a multitude of different activities that happen in computer networks, there are only few "elementary" data acquisition techniques, namely:

- Packet capturing (also known as sniffing);
- Connection tracking, and
- Netflow-based methods.

These techniques have different properties and different advantages and drawbacks. Packet capturing provides the most information [14] as, in principle, all transmitted information is accessible. However, the actual outcome depends on factors like, e.g., the placement of the data acquisition device, the network technology, or the network topology. In general, capturing (and analyzing) packets can be a costly operation. Considering a fully loaded Gigabit

 ISBN: 978-1-61208-279-0

Ethernet (1000Base-T) [15] transmitting Ethernet packets with a payload of 1,500 bytes only, we obtain a packet rate of approximately 82,000 packets per second; for minimum size packet (520 bytes) approximately 240,000 packets per second.

These rates are important because the monitoring system must be dimensioned accordingly. If the number of arriving packets exceeds the number of processed packets for a duration, the queue memory will eventually be exhausted and packets will be randomly discarded leading to information loss. Various approaches for achieving high capture rates exist, such as, special setups [8], optimized software [16], special hardware, or combinations thereof.

Connection tracking observes state changes of connections in a network [17]. While, formally, connection tracking is only applicable for stateful connection protocols such as, e.g., TCP [18], the notion of an implicit connection and states has also evolved for connectionless protocols, such as, e.g., UDP [19].

One advantage of connection tracking is that most implementations are highly optimized and known to perform very well. An example is nf_conntrack from the Linux Netfilter Project, which is, e. g., utilized by stateful Linux packet filters, like the well established iptables implementation. However, the amount of information which can be obtained from connection tracking is quite limited compared to packet capturing.

Netflows are also a performance-friendly form of providing information about network traffic in an aggregated form. A netflow is defined as unidirectional data flow between two endpoints and fully characterized by the 5-tuple of source and destination network layer address, source and destination transport layer port, and transport layer protocol type. When netflows are used for monitoring, further information about each flow is collected such as, e.g., packet counts and byte counts. From the perspective taken here, netflows can be considered similar to connection tracking and have the same advantages and disadvantages. It is worth mentioning that many commercial network interconnection devices likes switches and routers allow to export netflow data out of the box, and standards have been defined for exporting and importing netflow information [20][21].

A. Desirable Properties

In the following, we will outline desirable properties for a network analysis and surveillance system.

Distributed data collection: Nowadays, IT infrastructures and the underlying networks are spread topologically and topographically. Companies and governmental institutions have several subsidiaries, which are all integrated into the overall IT and network infrastructure. Even in smaller businesses the company network is usually composed of different subnetworks. In order to get a comprehensive overall picture of the situation in such networks, it is insufficient to collect data at a single place. Instead, data about all topographical and topological parts of the network is required. Hence, data must be gathered at a large number of different places in the computer network. The underlying network analysis and surveillance infrastructure must support this.

Heterogeneous data sources: many different ways for collecting data about computer networks exist. The individual approaches have advantages and disadvantages. While, e.g., packet capturing provides a very high detail of information, it is perfomance-intensive. Connection tracking, on the other hand, requires less performance but also offers less detailed information. The required level of detail strongly depends on the actual application scenario. Furthermore, as we will show, it is even possible to derive the same higher-level information from different types of natural events. Additionally, some approaches may deliver information that cannot be retrieved by other means. Thus, in order to acquire a complete overview of a network and the occurrences in it, ideally, multiple, heterogeneous data sources should be used.

Near realtime information: In computer networks, it is often required to swiftly respond to hazardous situations; e.,g., failures must be compensated for quickly or attacks must be stopped early. Therefore, the available information about a network must be as current as possible. Ideally, the information should be available immediately when incidents in a network happen. A network analysis and surveillance infrastructure should support the propagation of information in a timely manner such that the information is available in near realtime, i.e., “soon enough” or “in a timely manner” to react meaningfully [23]. We refer to this property as "near realtime" since strict definitions for “realtime” exist in other areas of computer science, particularly in the scope of embedded devices and realtime hardware.

Focused user interface: It is equally important that the important information can easily be found and that it is easily understandable. In the field of network analysis and surveillance, vast amounts of data are available. Thus, identifying the important information is crucial in order to enable taking correct decisions. Administrative personal can only perform its tasks efficiently and solve problems quickly if the important information is presented in an easily perceivable and understandable fashion. Confronting the administrative personal with too much information significantly slows down the decision making process. Presenting to much unimportant information may result in in-effective actions being taken. In the worst case, unimportant or even misleading information may result in the wrong, counter-productive decisions. Thus, a network analysis and surveillance system should assist the responsible administrative personal by presenting the important data in a way that is easily understandable and quickly perceivable.

While the above list of desirable properties of a network analysis and surveillance system is not meant to be extensive, we believe it covers the most important aspects for such a system. In the following, we will motivate our choices and results presented here by referring back to these desirable properties.

 ISBN: 978-1-61208-279-0

IV. COMPLEX EVENT PROCESSING

In the following, we will present the salient features of Complex Event Processing (CEP) and define some terminology. Please refer to the Event Processing Glossary [24] for more details.

Complex Event Processing (CEP) is an approach for processing data in form of events [1]. A typical usage scenario is to infer increasingly complex information from simpler information by correlating the simpler information with each other. Other frequently used actions are filtering or transformation of information.

In CEP, the basic unit of information is an "event". "Event", thereby, has at least two meanings. Firstly, so called "natural events" are occurrences that naturally happen in the field a CEP system is applied in. The field in which a CEP system is applied in is also referred to as the "application domain" of the system. In the application domain of network analysis and surveillance, a natural event is, e. g., a packet being transmitted or a connection state being changed. Secondly, the other meaning of "event" is the entity that is processed inside a CEP system and which contains information describing this event. This entity can also be seen as an "event object" that carries information inside a CEP system, possibly about a natural event. Please note that event objects inside a CEP system do not necessarily relate to natural events outside the CEP system. The information contained in such an event object is also referred to as "event properties".

Inside a CEP system, events are processed. One of the most powerful operations is the derivation of events from other events. We use the terms "simpler events" and "complexer events" to clarify the relationship between the processed and the derived events [25]. Complex events may be derived from simple events directly, but can be also be derived from other (intermediate) complex events.

The actual correlation of events happens in a component known as the "CEP engine". The CEP engine is one of the key components of every CEP system. For inferring complexer events from simpler events special processing rules are used. These processing rules are usually expressed in an Event Processing Language (EPL). The actual form and syntax of an EPL depends on the respective implementation. One popular way to express EPLs is in a way similar to Structured Query Language (SQL), but with additional functionality for correlating events [26].

A more general term in this context is "derived event". "Derived events" are events which were processed by some mean. Thereby, it doesn't matter if events had only been filtered or transformed or had been inferred from other events with means of a CEP engine.

Closely related to event objects is the idea of an "event type" or "event class". In a nutshell, an "event type" roughly defines the structure of concrete event objects. However, please note that an "event class" is not a class in the object-oriented paradigm. An event type definition is much less strict than a class definition in object-oriented languages. When considering the application domain of network analysis and surveillance one event type is that of a packet being transmitted. A packet can be of many different protocol types like TCP, UDP, ICMP, or ARP [27][28]. Clearly, all these protocols contain very different information and so do the resulting event objects inside the CEP system. So, an event type in a CEP system can be more seen roughly similar to an eXtensible Markup Language (XML) Schema Definition (XSD) than a class in the object-oriented sense.

When events are derived from each other, typically, events of one class or type are derived from events of another type. The relationship between these complexer and simpler events can be represented in a directed graph with the event types being the vertices and the edges resembling the derivation relations. The direction of the edges indicates the derivation relation from simpler events to complexer events. Such a graph representation of events and their relations is also referred to as "event hierarchy".

The event processing itself is usually composed of a larger number of other components. When the events are processed in a CEP system they usually flow through what is also called "Event Processing Network" (EPN). An EPN can consist of different components like event filters, event transformers, CEP engines, or communication infrastructures. In Figure 1, an exemplary EPN is depicted. The feedback loop at the event pattern matching component indicates that increasingly complex events can be derived in an iterative manner.

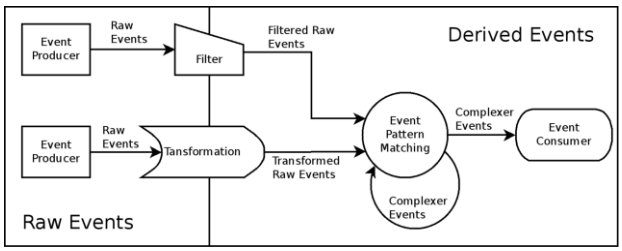


Figure 1: Simple Example EPN.

The communication infrastructure that is used in a CEP system can be manifold like common network sockets, higher-level infrastructures like Message-Oriented Middleware (MOM) [29] or an Enterprise Service Bus (ESB) [30] or traditional mechanisms for interprocess communication, like shared memory when components are locate on the same system.

V. OUR APPROACH

Much information about a computer network inherently has the nature of events. Many different things happen in computer networks all the time and naturally occur as events. The different events range from low-level to high-level. Examples of events happening in a network are packets being sent, connections changing the state, hosts "entering" or "leaving" a network, services being announced, requested, or used, devices failing, or attacks taking place.

Hence, in our opinion, an event-driven architecture [31] is the natural choice for processing data in the application domain of network analysis and surveillance [32]. In an event-driven system, the processing of data is triggered instantaneously upon the arrival of new data. Thus, an event-

 ISBN: 978-1-61208-279-0

driven system supports the processing of data in near realtime. This actively helps to fulfill the desired property that a network analysis and surveillance system shall operate in near realtime.

In event-driven architectures, data exchange also follows the event-driven paradigm. One approach for communicating data inside event-driven systems is via messages. Messages simply contain the event data that needs to be exchanged. Thereby, messages can be exchanged locally via "classical" interprocess communication or via different hosts across computer networks. This allows to integrate components from different spatially distributed locations. Thus, we can address the desired property of integrating distributed event sources this way.

However, using the event-driven paradigm alone is not sufficient for addressing the other desired property: a focused user interface that aids the administrative personal in identifying and solving critical situations. It is crucial that the important information is clearly and easily perceivable and understandable. Thus, a mechanism for efficiently filtering information and inferring high-level information is required. Consequently, we chose CEP as paradigm because this not only offers all features of an event-driven architecture but also allows to filter information and most important of it all enables to infer high-level information.

Furthermore, the message-driven event exchange results in a loose coupling between the individual components. The loosely-coupled infrastructure in combination with the CEP paradigm allows for the simple integration of heterogeneous components. In Figure 2, the general architecture of our proposed system is depicted. Our prototype takes advantage of all these properties. It combines the capability of integrating heterogeneous data sources at distributed locations with an EDA and a CEP engine.

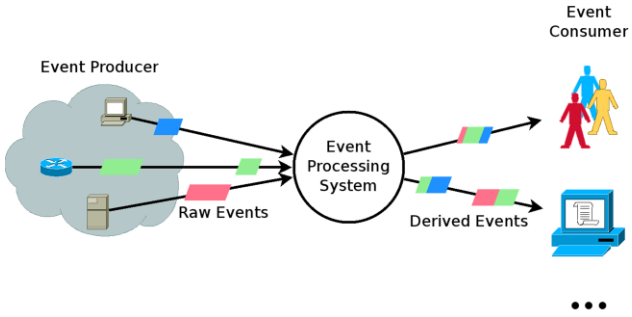


Figure 2: System Overview.

VI. EVALUATION

In order to assess the feasibility of our approach we implemented a CEP-based network analysis and surveillance system as prototype. We tested this prototype in real world network scenarios. Furthermore, we performed synthetic load tests for benchmarks.

In a CEP-based system, event patterns are used for processing information. For meaningfully applying a CEP system in a respective application domain, it is crucial that the intended application domain can be modeled with means of event patterns. To show that CEP can be meaningfully applied in the application domain of network analysis and surveillance we developed different event patterns and tested these in real-world network scenarios with our prototype. The event patterns were written in the Esper EPL language [26] and aim on detecting different situations in a computer network. These EPL patterns serve as proof of concept implementations and show that it is possible to effectively deduce meaningful high-level data in the application domain of network analysis and surveillance with means of CEP. In Figure 3, we show an exemplary event hierarchy for detecting Denial of Service (DoS) attacks.

We actually implemented different event patterns for the DoS attack scenario. It is, e. g., possible to conclude from a TCP SYN packet that is not followed by an established TCP connection that something irregular is happening; only the initial SYN packet was observed but the actual three-way handshake didn't happen. When there is a large number of such events within a short time it can be concluded that a SYN flood is going on and in turn it can be concluded that a DoS attack is performed. The act of checking for the absence of a certain event as described here is also referred as the "absence pattern" [33]. Similarly, it could be concluded that an DoS attack is going on by just looking at the frequency with which SYN packets are sent without looking for non established connections. In the event hierarchy shown in Figure 3 the optional ways for deriving events are indicated by dashed edges. For the sake of brevity we only depict and explain one exemplary event hierarchy. However, we could successfully implement many other patterns, e.g., for detecting ARP spoofing attack, congestion situations, or brute-force attacks. This example illustrates that the domain of network analysis and surveillance can be modeled with means of event patterns.

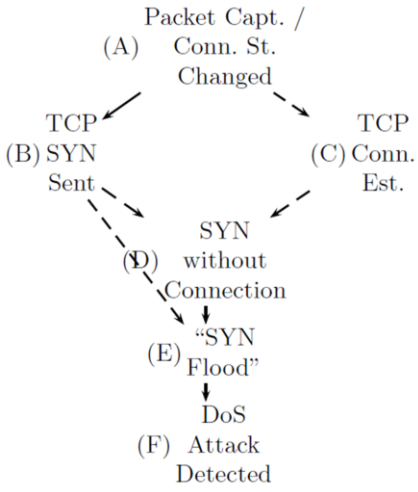


Figure 3: Exemplary Event Hierarchy for Detecting Denial of Service Attacks.

Furthermore, in the example from Figure 3, it is also possible to use events from connection tracking for detecting DoS attacks. This property can, e.g., be leveraged for optimizing the performance of a network analysis and surveillance system. In our prototype, we implemented and

 ISBN: 978-1-61208-279-0

tested the DoS detection pattern shown in Figure 3 for natural events from packet capturing and connection tracking. Both variants showed to be capable of detecting the same DoS situations. With this example, we could show that, thanks to the loose coupling enabled by the CEP and EDA paradigm, the integration of heterogeneous data sources with our approach is simple and quick.

In order to assess the performance of a CEP system in the application domain of network analysis and surveillance, we exemplary used the Java-based, general purpose Esper CEP implementation. For our benchmarks, we used three different event patterns from the targeted application domain based on natural events from packet capturing: a simple pattern for simply filtering the input data, a pattern with average complexity for calculating the time between the occurrence of certain packets, and a complex pattern for detecting the TCP three-way handshake. The sample data had been collected with Wireshark. For the benchmarks, the data was synthetically fed to the CEP engine at different rates. For each measurement, the rate with which the data was "replayed" was constant. During each measurement run, we determined the CPU usage of the system and the percentage of events that could be successfully processed. The computer system on which we made the benchmarks was a common consumer class laptop with an Intel Core i5 CPU and 4GB memory. In order to avoid measuring artifacts of multi-CPU features as offered by this CPU, like impacts on multi-threading, etc., we artificially disabled the multi-processor functionality of the CPU.

The results of our benchmarks are shown in Figure 3. Please note that we used the rate of input events for the x-axis. The overall rate of events inside the CEP engine is usually higher because the derived events also attribute to the overall event rate. Similarly, we calculated the ratio of successfully processed input events for the statistics.

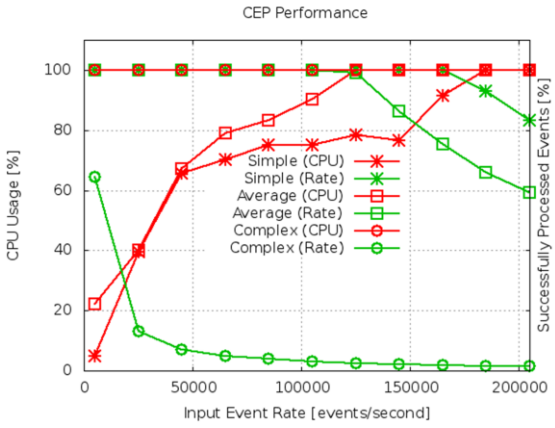


Figure 4: CEP Performance Comparison for Events with Different Complexity.

It can be clearly seen that the actual performance strongly depends on the complexity of the respective event patterns. Thereby, the percentage of successfully processed events is approximately 100% as long as enough CPU resources are available. Once the CPU resource limit is reached, the percentage of successfully processed events continuously decreases with increasing input event rate. Thus, we consider the highest rate of input events for which the percentage is still 100% as the maximum achievable rate for the respective event pattern. For the simplest patterns, we could achieve a throughput of ~170.000 events per second; with enabled multi-processing this was even higher, in the magnitude of 500.000 events per second. However, with multiple CPU cores, the result data showed some irregularities that we attribute to the multi-CPU features. For the sake of clarity we, therefore, disabled the multi-CPU support via the operating system. For the average complexity pattern, the maximum achievable throughput in our test setup is about 125.000 events per second. Meaningfully processing the most complex pattern was not possible at all; even at a rate of 5000 input events per second it was not possible to get a event processing rate near 100%. Still, this shows that even a general purpose CEP engine on consumer class commodity hardware CEP can generally perform with Gigabit Ethernet speed with a packet rate of roughly 82000 events per second. Yet, the actual performance strongly depends on the respective event patterns. With increasing event pattern complexity the performance degrades until it becomes impossible to meaningfully process the input data. However, this strongly depends on the applied event patterns. Furthermore, we purposely used packet capturing as data acquisition approach as it has the highest rate of emitting data. Other approaches like connection tracking or netflows emit data at a much lower speed. Additional ways for compensating performance issues are additional pre-processing and filtering steps.

Also, note that the CEP engine we used for benchmarking is a general-purpose implementation, which is intended to run on a large variety of platforms. In the field of CEP, there is also work and research on highly optimized CEP engines that leverage hardware acceleration.

VII. Conclusion and Future Work

Computer networks are crucial for the operation of nowadays IT infrastructures. Failures in computer networks very often directly impact the functionality of the corresponding IT with possibly severe consequences. Thus, maintaining operational computer networks is highly important.

Modern data processing paradigms, modern IT infrastructures and architectures, and increased performance open up new possibilities for gathering, processing, combining, and using data. We take advantage of this and propose an improved approach for network analysis and surveillance.

Based on an overview of existing approaches for network analysis and surveillance, we defined and explained desirable properties for a modern network analysis and surveillance system. Our approach on addressing these requirements is to leverage the CEP and EDA paradigm of data processing. Based on these technologies, we could successfully implement a distributed network analysis and surveillance system prototype that operates in near realtime and offers powerful functionality for processing, filtering, and enriching

 ISBN: 978-1-61208-279-0

information. We tested our system in real world networking scenarios and with benchmarks. The results, so far, show that our approach works and is suited to fulfill the requirements we stated.

In future, we are going to further extend our system. We are currently working on integrating machine learning and anomaly detection techniques into the system. This way, we will further improve the capability of the system for deducing meaningful information and detecting important situations. We will also test the system in more scenarios and will optimize the performance.

REFERENCES

[1] D. C. Luckham, "The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems," Addison-Wesley Longman Publishing Co., Inc., Amsterdam 2001.

[2] M. Ficco and L. Romano, "A Generic Intrusion Detection and Diagnoser System Based on Complex Event Processing," First International Conference on Data Compression, Communications and Processing (CCP), IEEE Computer Society, Los Alamitos, 2011, pp. 275-284.

[3] J.J. Martinez-Molina, M.A. Hernandez-Ruiz, M.G. Perez, G.M. Perez, and A.F. Gomez-Skarmeta, "Event-Driven Architecture for Intrusion Detection Systems Based on Patterns," Second International Conference on Emerging Security Information, Systems and Technologies SECURWARE '08, Cap Esterel, 2008, pp. 391-396.

[4] L. Aniello, G.A. Di-Luna, G. Lodi, and R. Baldoni, "A Collaborative Event Processing System for Protection of Critical Infrastructures from Cyber Attacks," Proceedings of the 30th International Conference on Computer Safety, Reliability, and Security SAFECOMP'11, Springer-Verlag Berlin, Heidelberg, 2011, pp. 310-323.

[5] L. Aniello, G. Lodi, and R. Baldoni, "Inter-domain Stealthy Port Scan Detection through Complex Event Processing," Proceedings of the 13th European Workshop on Dependable Computing EWDC '11, ACM New York, 2011, pp. 67-72.

[6] V. Krishnamoorthy, N.K. Unni, and V. Niranjan, "Event-driven Service-oriented Architecture for an Agile and Scalable Network Management System," International Conference on Next Generation Web Services Practices, 2005.

[7] S. Cheng, Z. Cheng, Z. Luan, and D. Qian, "NEPnet: A scalable monitoring system for anomaly detection of network service," 7th International Conference on Network and Service Management (CNSM), 2011.

[8] C. Morariu and B. Stiller, "DiCAP: Distributed Packet Capturing architecture for high-speed network links," 33rd IEEE Conference on Local Computer Networks, LCN 2008.

[9] H. Inoue, T. Takenaka, M. Motomura, "20Gbps C-Based Complex Event Processing," International Conference on Field Programmable Logic and Applications (FPL), 2011.

[10] Nagios Enterprises, "Nagios - The Industry Standard in IT Infrastructure Monitoring," Online, http://www.nagios.org/, last accessed 2013-02-16.

[11] Zabbix SIA, "Homepage of Zabbix, An Enterprise-Class Open Source Distributed Monitoring Solution," Online, http://www.zabbix.com/, last accessed 2013-02-16.

[12] S.A. Shaikh, H. Chivers, P. Nobles, J. A. Clark, and H. Chen, "Network Reconnaissance," Network Security, 2008.

[13] S. Webster, R. Lippmann, and M. Zissman, Experience Using Active and Passive Mapping for Network Situational Awareness, Fifth IEEE International Symposium on Network Computing and Applications, 2006.

[14] C. Sanders, "Practical Packet Analysis : Using Wireshark to Solve Real-World Network Problems," No Starch Press, Incorporated, San Francisco, 2007.

[15] The Institute of Electrical and Electronics Engineers, Inc., "IEEE Std 802.3-2008 Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications," 2008.

[16] L. Deri, "nCap: wire-speed packet capture and transmission," *Workshop on End-to-End Monitoring Techniques and Services, 2005*, pp. 47-55.

[17] P. Ayuso, "Netfilter's Connection Tracking System," LOGIN: The USENIX magazine, Berkeley, 2006.

[18] J. Postel, "Transmission Control Protocol – RFC 793 (Standard)," Request for Comments, Internet Engineering Task Force, 1981.

[19] J. Postel, "User Datagram Protocol – RFC 768 (Standard)," Request for Comments, Internet Engineering Task Force, 1980.

[20] P. Phaal, S. Panchen, and N. McKee, "InMon Corporation's sFlow: A Method for Monitoring Traffic in Switched and Routed Networks – RFC 3176," Request for Comments, Internet Engineering Task Force, 2001.

[21] B. Claise, "Cisco Systems NetFlow Services Export Version 9 – RFC 3954," Request for Comments, Internet Engineering Task Force, 2004.

[22] M. Natu and A.S. Sethi, "Active Probing Approach for Fault Localization in Computer Networks," 4th IEEE/IFIP Workshop on End-to-End Monitoring Techniques and Services, 2006.

[23] C. Ballard, M.S. Roopa, O. Mueller, Z.Y. Pen, A. Perkins, and P.J. Suresh, "Preparing for DB2 Near-Realtime Business Intelligence," IBM Redbooks, 2004.

[24] D.C. Luckham, et al., "Event Processing Glossary – Version 2.0," Online, http://www.complexevents.com/wp-content/uploads/2011/08/EPTS Event Processing Glossary v2.pdf, last accessed 2013-02-16.

[25] R. Gad, J. Boubeta-Puig, M. Kappes, I. Medina-Bulo, "Hierarchical Events for Efficient Distributed Network Analysis and Surveillance," Proceedings of the 2nd International Workshop on Adaptive Services for the Future Internet, WAS4FI-Mashups '12, ACM New York, 2012, pp. 5-11.

[26] EsperTech Inc., "Esper - Event Stream and Complex Event Processing for Java Reference Documentation," Online, http://esper.codehaus.org/esper/documentation/documentation.html, last accessed 2013-05-02.

[27] J. Postel, "Internet Control Message Protocol – RFC 792 (Standard)," Request for Comments, Internet Engineering Task Force, 1981.

[28] D.C. Plummer, "An Ethernet Address Resolution Protocol or Converting Network Protocol Addresses – RFC 826 (Standard)," Request for Comments, Internet Engineering Task Force, 1982.

[29] B. Snyder, D. Bosanac, and R. Davies, "ActiveMQ in Action," Manning Publications, 2011.

[30] D. Bo, D. Kun, and Z. Xiaoyi, "A High Performance Enterprise Service Bus Platform for Complex Event Processing," Seventh International Conference on Grid and Cooperative Computing, IEEE Computer Society Washington, DC, 2008, pp. 577-582.

[31] H. Taylor, A. Yochem, L. Phillips, and F. Martinez, "Event-Driven Architecture: How SOA Enables the Real-Time Enterprise," Addison-Wesley Professional, Boston, 2009.

[32] R. Gad, J. Boubeta-Puig, M. Kappes, and I. Medina-Bulo, "Leveraging EDA and CEP for Integrating Low-Level Network Analysis Methods into Modern, Distributed IT Architectures," VII Jornadas de Ciencia e Ingeniería de Servicios (JCIS - SISTEDES 2012), Almería, Spain, 2012.

[33] O. Etzion and P. Niblett, "Event Processing in Action," Manning Publications Co., 2010.

 ISBN: 978-1-61208-279-0

MYHand: a Novel Architecture for Improving Handovers in NGNs

Mario Ezequiel Augusto
Department of Information Systems
Santa Catarina State University, Brazil
Email: mario.augusto@udesc.br

Renata Porto Vanni
Federal Institute of Sao Paulo
Campus Araraquara, Brazil
Email: rportovanni@ifsp.edu.br

Helio Crestana Guardia
Department of Computer Science
Federal University of Sao Carlos, Brazil
Email: helio@dc.ufscar.br

Mahdi Aiash
School of Science and Technology
Middlesex University, London
Email: M.Aiash@mdx.ac.uk

Glenford Mapp
School of Science and Technology
Middlesex University, London
Email: G.Mapp@mdx.ac.uk

Edson dos Santos Moreira
Department of Computer Systems
University of Sao Paulo, Brazil
Email: edson@icmc.usp.br

***Abstract*—The on demand access, provided by Next Generation Networks (NGN), will allow users of mobile devices to choose from and connect to networks with no pre-established service contract. Besides signal strength, knowledge about different parameters of the available networks shall base the selection of the attachment point to use. No mechanism currently available provides the desired integrated support for network discovery and on demand access. This paper presents MYHand, an architecture for providing extended information in NGN scenarios. By using the IEEE 802.21 protocol Basic Schema and part of the Y-Comm architecture, MYHand improves the handover managed by mobile devices (user-centric management). This paper also presents an extension to the IEEE 802.21 Basic Schema, which is used by MYHand for extra information exchange between mobile devices and heterogeneous networks.**

Keywords-NGN; MIH; Y-Comm; Handover Management

I. INTRODUCTION

Internet service providers currently share a relatively static market where a multitude of mobile devices send and receive data using different wireless technologies. Costs and different aspects of Quality of Service (QoS) are key factors for customer fidelity. Next Generation Networks (NGN) [1] will change this scenario by providing support for multimedia services and device mobility, accompanied by mechanisms for network discovery and selection. Other features of NGNs include the simultaneous support for different transmission technologies and overlapping network coverage.

NGNs will bring the user to the center of a handover decision process, which shall be done transparently, by matching pre-established desired QoS parameters with the characteristics of the available networks. Handover is defined as the switching of the Point of Attachment (PoA) of a mobile device [2]. A handover can be classified as a horizontal handover, which occurs when the new point of attachment is technologically identical to the previous one, or as a vertical handover, which occurs when the new PoA is technologically different to the previous one [3].

An advanced classification of handover divides it into two types: imperative or alternative [4]. Imperative handovers occur due to technological reasons such as signal strength, coverage and QoS, and it is called imperative because there may be a severe loss of performance if they are not performed. Alternative handovers occur due to reasons other than technical issues, such as, pricing, incentives, preferences, context, or available services.

Information about networks within reach includes SSIDs, signal strength and noise ratio, when using IEEE 802.11 interfaces, and network IDs, and frequency related parameters for different 3G networks. Network discovery, however, may imply switching different communication interfaces into a costly scanning procedure. Besides, several QoS parameters, and dynamic billing information cannot be observed by such mechanisms.

Authentication, authorization and accounting (AAA) also challenge the viability of NGNs since, in the envisioned on-demand service model for NGNs, no fixed contract will be required to allow an user access to the available network infrastructures. The use of multiple access technologies becomes an implication, because different terms can be used for the same information. For instance, parameter jitter is called ”jitter” in IEEE 802.11, ”tolerated jitter” in IEEE 802.16, and ”delay variation” in 3GPP networks.

No mechanism currently available provides the desired integrated support for network discovery and on demand access. Even if the support for IEEE 802.21 [5] services are available in a network, terminology issues make it difficult to correctly detect events and match the desired QoS with the offered services. Only a rich and coherent set of information will enable the envisioned dynamic and on demand service selection in NGNs. Dynamic handovers and the free competition among providers shall then benefit users in a virtuous cycle.

As the variety of wireless technologies and mobile devices is increasing, the discovery and selection of networks is becoming an important issue. This paper presents MYHand, an architecture for providing the mobile devices with additional information for dynamic handover decisions in Next Generation Networks. The name MYHand stands for ”MIH-based

 ISBN: 978-1-61208-279-0

and Y-Comm-based Handover Management". In the MYHand architecture, network information is provided to the nodes via Y-Comm [4] [6], along with instances of the information service (MIIS), events service (MIES), and command services (MICS) of the IEEE 802.21 protocol [5]. An extension to the IEEE 802.21 Basic Schema (Extended Schema) was also introduced wherewith the provider can offer additional information to the mobile devices, including incentives, thus increasing competition among access providers. MYHand optimizes the handover process as it aids in the early and effective discovery of available networks. A flowchart is presented, which details an alternative handover procedure with minimum throughput requirement, using MYHand architecture. Simulation results based on this architecture show that the mobile user could prioritize a preference without loosing the access quality.

Unlike other works, the proposed extended schema focuses on alternative handover, although MYHand could also be used in imperative handovers. In addition, MYHand is user-centric, i. e., the handover if managed by the user device, as opposed to network-centric, offering greater freedom of choice.

As contributions of this work, we highlight the benefits for providers and mobile users. By adopting the proposed architecture, providers can disclose additional informations other than usual, for instance, incentive information, and thus attract new users. Besides, MYHand can be extended to offer any kind of information, including service offering. The information provided by the MYHand architecture will help the mobile in the discovery and selection of an access network.

The rest of the paper is organized as it follows. Section II presents some approaches to obtain information about available networks. Section III presents some related works. MYHand architecture is presented in Section IV. The last section concludes the paper and suggests some future work.

II. STRATEGIES FOR SELECTING TRANSMISSION INFRASTRUCTURE IN NGN

The choice of a network depends on the knowledge of the available options at each time. For this purpose, an operational entity running on the device in the form of a high-level process or something embedded inside the kernel should do a matching between the user desired features in terms of price and QoS parameters, for instance, and the available options.

There are different approaches to find out the available networks in the mobile device vicinity. At a lower level, it is possible to scan for the available access points of each network interface on the device. Using IEEE 802.11 networks, for instance, it is to know the available access points (APs) and their characteristics such as frequencies and signal strength. The same goes to Wimax and LTE networks. To this end, the decision-making entity should start a scanning process with the desired frequency. A consequence of this periodic scanning is the transmission interruption by the interface being queried, thus decreasing the throughput and increasing the power consumption.

The IEEE 802.21 protocol [5] introduces events that could minimize the need for periodic scanning for mapping available

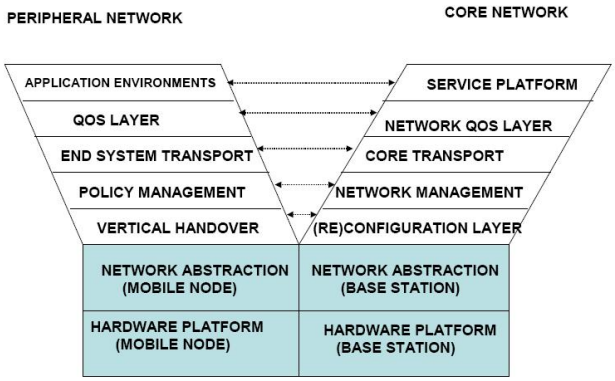


Fig. 1. Y-Comm Architecture, extracted from [6]

access points. If supported by a device, this standard foresees that the network interface itself, possibly using driver support in the operating system, performs a search for the desired network and generate notifications for a client entity that registers interest in such information. At a broader level, an element called MIIS (Media Independent Information Service) may be present in some device on the same network to collect information on the available access points and provide them later to probing customers. In this case, it avoids scanning on each mobile device. The communication between the client decision entities and the MIIS server occurs using application protocols over TCP/IP.

In the Y-Comm project [4] [6], an entity present in the Network Management layer performs equivalent functions to those provided by the MIIS, sending some information to the mobile, such as, network topology, resources, QoS parameters, etc. Obtaining such information also minimizes the need of scanning. The possibility of performing authentication on demand with a target network extends the functionality provided by the IEEE 802.21, but it is expected in Y-Comm, which communicates with several access providers.

The Y-Comm architecture is divided into layers and it is composed by two frameworks, as shown in Fig. 1. The Peripheral Framework, implemented in the Peripheral Network, deals with operations and functions on the mobile, and the Core Framework, implemented in the Core Network, which deals with the functionality required in the core network to support the Peripheral Framework. A detailed explanation about each layer can be found in [4] and [6]. In this paper, the terms Peripheral Framework and Peripheral Network are used interchangeably (the same to Core).

To understand how Y-Comm does handover, Fig. 2 shows a proactive handover procedure. The Network Management layer (NML) provides the AAA system, which is not provided by IEEE 802.21, and stores information about local networks. In the mobile, the Policy Management layer (PML) polls the Network Management layer (NML) to obtain information about all local wireless networks, their topologies and QoS characteristics. This information, along with others provided by higher layers such as, location, speed and direction are

 ISBN: 978-1-61208-279-0

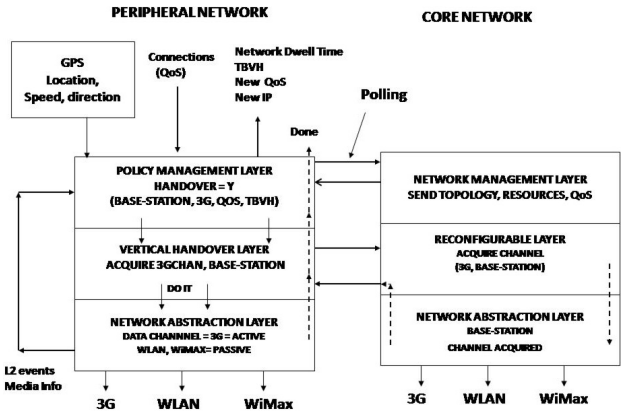


Fig. 2. Proactive handover using Y-Comm, extracted from [4]

used by the PML to evaluate the circumstances under which a handover should occur. The PML can be configured with some rules related to handover decision. The PML also calculates the Time Before Vertical Handover (TBVH) and communicates this information to the Vertical Handover layer (VHL), which requests resources to the Reconfiguration layer. In addition, once the PML decides to hand over, the new IP address, the new QoS, the TBVH and estimated Network Dwell Time are communicated to the upper layers. The Network Abstraction layer, both in the peripheral and core network, is responsible for providing a common interface to manage and control all the network links, and for sending L2 events to the PML.

MYHand is an application of the dynamic negotiation model trading under Y-Comm project, adding facilities of the MIHF (Media Independent Handover Function) services (event, information and command) supported by an extension of the Basic Schema. This extension provides a more efficient interaction between mobile and access network in terms of information exchange.

III. Related Work

There are several works related to network discovery and handover optimization using IEEE 802.21 protocol. In [7] the authors propose a multiple attributes decision making-based terminal controlled vertical handover decision schema using IEEE 802.21. The proposed schema is compared to RSS-based and cost function-based schema through simulations, which show that the proposed schema provides smaller handover times and lower dropping rate than the RSS-based and cost function-based vertical handover methods. But the authors focus on decision making in the integrated Wi-Fi and Wimax networks. MYHand architecture is designed to work with any network technology. In [8] the authors propose an architecture of MIIS server and the procedures for handover optimization, which avoid scanning and reduces energy consumption, but the management is network-controlled. Both works are aimed for imperative handover.

A lot of work focuses on handover optimization. Rizvi et. al. [9] presents an intelligent vertical handover decision algorithm and points out other works. In [10] the authors present an overview of the handover decision strategies, which are classified in categories, and present a new approach which considers context-aware and policies, aided by a Fuzzy Logic system. In [11] the authors present a multiservice vertical handover decision algorithm (MUSE-VDA) and a general cost function used to choose a target network. In [12] the authors describe a policy based handover decision algorithm (POLIMAND) and point several link layer parameters in heterogeneous networks used in the decision making. These works do not use IEEE 802.21, in a way that higher layers do not receive notifications from the lower layers when an important event occur.

The IEEE 802.21 protocol does not specify how the information of the available networks is filled in the databases. The authors in [13] propose a mechanism for this. Thus, this related work can be used as an adjunct to MYHand. In [14] the authors propose a new architecture for network discovery and a solution for the construction of the information database. Their work focuses in mobile-assisted and network-assisted proactive handover (when the mobile attempts to know the condition of the various networks at a specific location before the mobile node reaches that location) and pre-authentication. In this architecture, the information stored in the servers is restricted to the ones registered by the mobiles from visited networks and the information that Reporting Agents (RAs), present in each network, catch via SNMP and send it to the information server. In [15] the authors propose an Hierarchical IEEE 802.21 Information Service Management infrastructure, which places MIIS servers in three levels: Zone MIIS, Local MIIS, and Global MIIS, aiming to improve the response time. In the MYHand architecture, the Y-Comm information server obtains information from many different places as, for instance, its local database, MIIS servers, and other information services such as the WFP server [16].

The works related to network discovery found in the literature concerned with technological information, needed for imperative handover process. MYHand extends the network discovery by embedding additional information related, for instance, to incentives, required for alternative handover process [17]. The extended schema presented in this paper is focused on alternative handover, but MYHand can optimize both, imperative and alternative handovers.

IV. MYHand architecture

This section presents the extended schema and the MYHand architecture, as well as its validation.

A. IEEE 802.21 Information Service Schema

The Information Service Schema is an RDF/OWL ontology (Resource Description Framework / Web Ontology Language). The schema is used in the IEEE 802.21 Information Service to define the structure of each information element, as well as the relationship among the information elements. The IEEE 802.21 Information Service schema is supported by every MIHF that implements the MIIS to support flexible and efficient information queries.

 ISBN: 978-1-61208-279-0

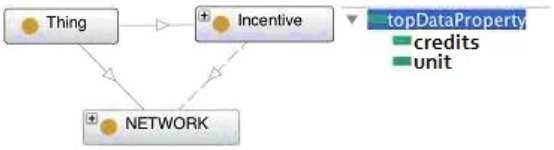


Fig. 3. Proposed Extended Schema Relationship

The RDF/OWL schema definition for MIIS consists of two parts: the basic and the extended schema. The MIIS RDF/OWL representation method is extensible using an extended schema.

B. Proposed Extended Schema

Due to the independence of service-related functions from underlying transport technologies in NGNs, and appearance of new technologies, it is expected more concurrence among access providers. The proposed schema extends the selection of new networks, embedding information related to incentives, enriching the alternative handover, in which the device may choose the target network based on incentives.

Fig. 3 shows the proposed extended schema in the form of a new element, called Incentive, which is related to Thing and NETWORK. Thing is a generic element of RDF/ OWL language which defines the basic type for an element, from which all other elements inherit their properties. The element NETWORK aggregates information from other elements, related to a certain network, such as network type, Point of Attachment list, Operator ID, among others, and the Incentive element becomes another attribute of NETWORK.

The element Incentive is a simple example on how MYHand architecture can improve handovers. This element provides information about amount of credits offered to the user, who connects to a certain access provider. This amount of credits can be used for future connections, for gaining discounts for example, and thus be an attractive to the loyalty of the user. This element has two properties, according to Fig. 3:

Credits: credits to be assigned to the user during the network usage. It is associated to property 'unit';

Unit: amount of time, in seconds, that the user must keep connected for gaining that amount of credits. For instance, the incentive is 3 credits for each 60 seconds connected.

Part of the proposed extended schema definition is shown in Fig. 4 as an XML document. The extended schema is obtained through DHCP service by the same way that the Basic Schema [5].

```
<owl:Class rdf:about="file:&mihextended;Incentive"/>

<owl:ObjectProperty rdf:about="file:&mihextended;incentive">
   <rdfs:range rdf:resource="file:&mihbasic;NETWORK"/>
   <rdfs:domain rdf:resource="file:&mihextended;Incentive"/>
</owl:ObjectProperty>

<owl:DatatypeProperty rdf:about="file:&mihextended;credits">
   <rdfs:comment>Credits to be assigned to user during the network
     usage. It is associated to property 'unit'.</rdfs:comment>
   <rdfs:range rdf:resource="&xsd;integer"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:about="file:&mihextended;unit">
   <rdfs:comment>Amount of time, in seconds, that the user must keep
     connected for gaining the amount of credits. For instance, the incentive
     is 3 credits for each 60 seconds connected.</rdfs:comment>
<rdfs:range rdf:resource="&xsd;double"/>

</owl:DatatypeProperty>
```

Fig. 4. Proposed extended schema definition

C. MYHand architecture

Fig. 5 presents the MYHand architecture with four entities involved: the MobileNode (MN), the Serving Point of Attachment (PoA-S), with a co-located Serving Point of Service (PoS-S), a Candidate / Target Network and a MIIS Server, each of them with IEEE 802.21 modules properly inserted in the Y-Comm layers.

The Mobile Node (MN) can have more than one network interface. The MIHF module is located in the Network Abstraction Layer and it is responsible for the abstraction of the network interfaces to the higher layers. This module receives commands and sends events and information to the Handover Manager module, receives commands from Vertical Handover Manager and forwards commands from higher modules to remote MIHFs. The Service Access Points MIH_SAP, MIH_LINK_SAP and MIH_NET_SAP are the interfaces between the MIHF and the other modules. The Handover Manager module acquires information from the device, user and networks (MIH_Get_Information and MIH_MN_HO_Candidate_Query commands) and it decides when and to which antenna a handover should be done. The Vertical Handover Manager module receives commands from the Handover Manager module calling for a handover (do_handover), acquires target network resources and actually performs the handover (MIH_MN_HO_Commit). Both, the Handover Manager and Vertical Handover Manager modules are implemented as MIH Users. The Mobile IP protocol allows the user mobility at network level.

Still in Fig. 5, the PoA-S/PoS-S is the network point to which the MN is directly connected. In the PoS-S and other PoSs, the MIH-LINK-SAP is not necessary because the MIH Users do not need to communicate with modules below the MIHF. The MIHF module abstracts the network interface to the upper layers and forwards remote MIH commands to the respective modules. The Handover Manager module acts as an MIIS Server proxy and it is the responsible for providing information to the MN. This information can be gathered from a local database, a server information MIIS Server (MIH_Get_Information) or other possible sources. This module also verifies resource availability at candidate networks by means of the MIH_N2N_HO_Query_Resources message. The Resources Manager module (RM) requests resource allocation to the MN in the target network.

In the Candidate/Target Network, the MIHF module also abstracts the network interface to the upper layers and forwards

 ISBN: 978-1-61208-279-0

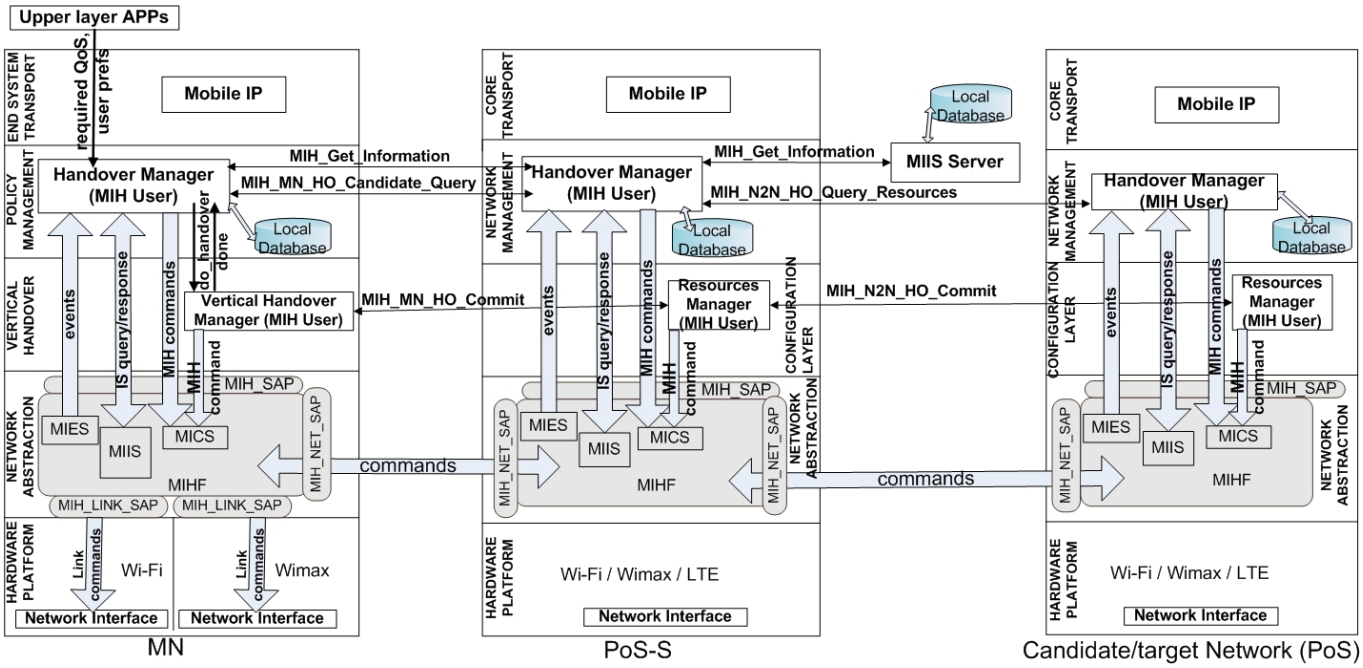


Fig. 5. MYHand architecture

remote MIH commands to the respective modules. The Handover Manager module checks and informs to the Handover Manager of the PoS-S about resource availability for new connections. The Resources Manager module (RM) allocates network resources to an user who will connect to this access point, when requested by the RM of the PoS-S.

In the MIIS Server, the module which implements the information server receives and responds to information requests about available networks in a given area.

For better understanding the architecture operation, Fig. 6 shows a flowchart which details an alternative handover procedure with minimum throughput requirement, which prioritizes the amount of credits that a provider offers, as explained in Section IV-B. IEEE 802.21 commands are started by ”MIH_”.

Before starting the handover, the MN is connected to network A by means of its Wi-Fi interface and it is running an application with a minimum throughput requirement. Higher layer applications in MN inform the Handover Manager module (with an appropriate frequency) on information about minimum QoS and user preferences (provider, access technology, price or minimum credits, for instance). In the case the MN is not yet connected to any network, a scanning is necessary, through the MIH_Link_Actions.request command, to search for an available network. This command reports information about AP address, network ID, signal strength, among others.

To enable an alternative handover, the Handover Manager module (HM) of the mobile node periodically requests information to the HM of the PoS-S on available networks in that area by means of the MIH_Get_Information.request message. In this message, the MN sends its location and desirable access technology (according to its available network interfaces and user preferences) and receives information (MIH_Get_Information.response) on each network as, for instance, provider SSID, roaming partners, access cost, maximum data rate, and credits to be gained (in the case of our extended ontology), and information about each available antenna as, for instance, MAC address, technology, geographic location, channel, and IP address. Eventually the HM of PoS-S can query the MIIS to update information. All remote MIH messages pass through the MIHF of the respective entities.

The MN can identify each network as belonging to a provider to which it has a contract or belonging to a partner of its home provider (through the roaming partners information of the IEEE 802.21 protocol).

If, after a query for available networks, the MN identifies a network offering more credits than the network currently being used, the HM module can decide to do a handover to that network. As the application requires a minimum throughput, the HM module sends an MIH_MN_HO_ CandidateQuery.request message to the HM module of the PoS-S, requesting resources verification, stating which networks must be checked (candidate networks) and the minimal required resources. This module, in turn, queries the resources availability in each candidate network by sending an MIH_N2N_HO_QueryResources.request message to the HM module of each candidate network, advising which resources are required. So, the HM of the PoS-S joins the information about those networks and responds to the MN with an MIH_MN_HO_CandidateQuery.response message.

Having all the necessary information, the HM module of the MN decides if a handover should be done. If so, it sends the do_handover command to the Vertical Handover Manager (VHM) to proceed with the handover, stating the link type (Wi-Fi, LTE, etc) and to which network the handover should be

 ISBN: 978-1-61208-279-0

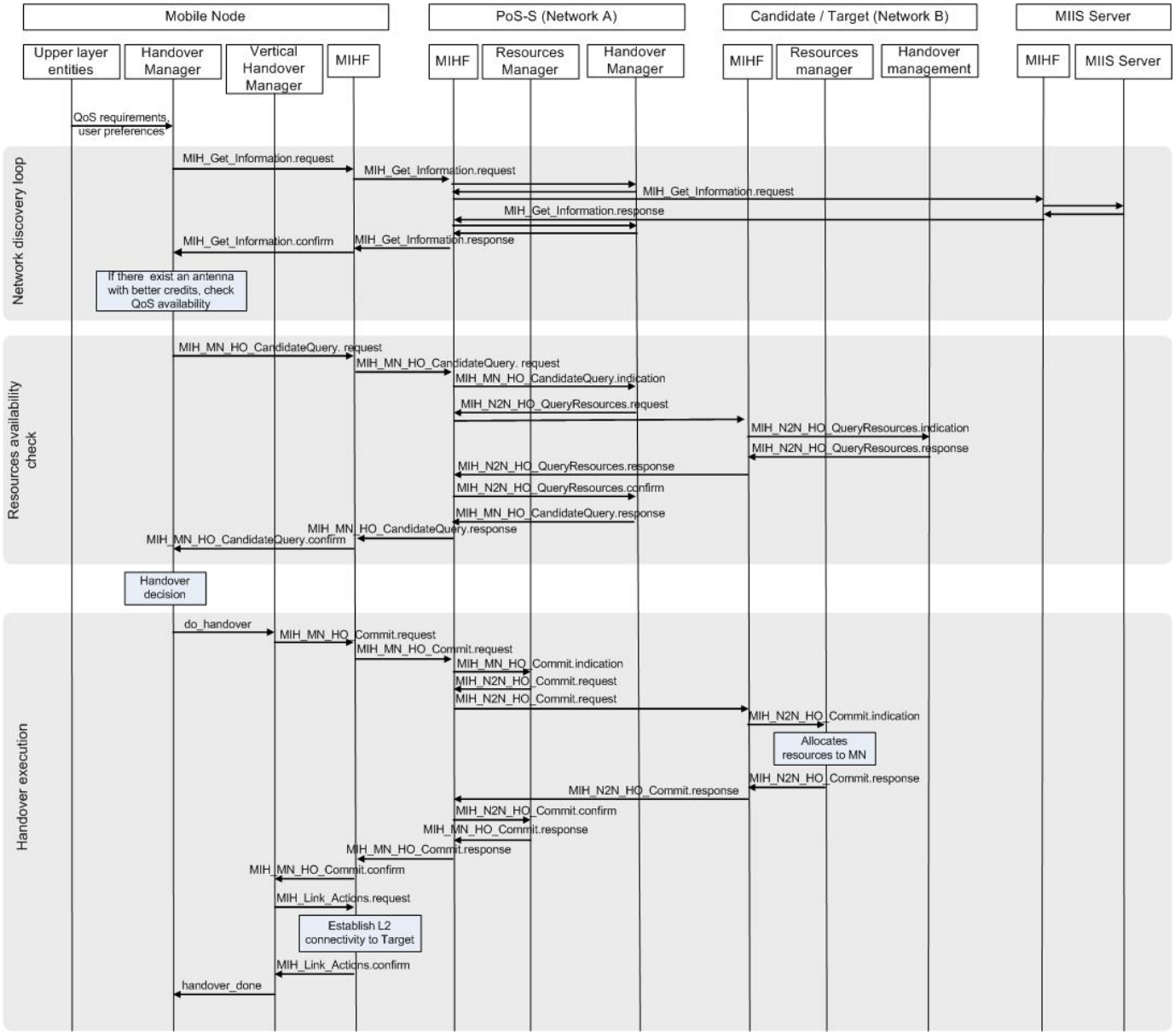


Fig. 6. Alternative handover flowchart

done. The VHM module notifies the Resources Manager (RM) of the PoS-S about the selected target network, by means of an MIH_MN_HO_Commit.request message. The RM module of the PoS-S sends the MIH_N2N_HO_Commit.request message to the RM module of the target network to advise that the mobile will connect to that network. Then the RM of the target network allocates the resources for the MN and replies sending a MIH_N2N_HO_Commit.response message advising on sucess or not. If so, the RM module of the PoS-S notifies the VHM of the mobile on the success of the operation by means of an MIH_MN_HO_Commit.response message.

Having the resources already allocated on the target network, the VHM module sends an MIH_Link_Actions.request command to the MIHF requesting that the interface is turned on to establish the connection at the link level to the target antenna in a given channel, or requesting that the current interface disconnects the current antenna and connects to the target antenna. Finally, the connection to the target is established and the VHM module informs the HM module that the handover is done (handover_done message).

After, Mobile IP is executed in the MN, Home Agent and Foreign Agent to keep the connection at transport layer.

D. Architecture Modelling Validation

For validating the modelling of the MYHand architecture, a scenario with 3 access providers (P1, P2 and P3) was simulated by using NS2 (Network Simulator 2) [18]. Although NGN networks foresees the usage of different technologies with signal overlay, the aim of this validation is focusing in

 ISBN: 978-1-61208-279-0

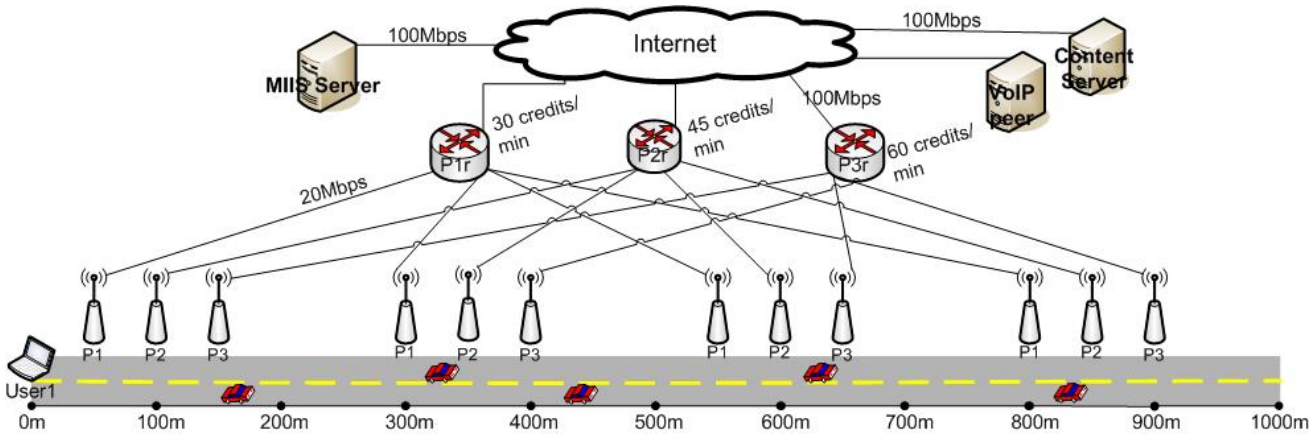


Fig. 7. Simulated scenario

the MYHand architecture with information about incentive. Because of this, only one wireless technology was simulated. Each provider has 4 Wi-Fi antennas positioned along an 1000 meters avenue, as showed in Fig. 7. Each antenna covers, approximately, 200 meters diameter. P1r, P2r and P3r are edge routers belonging to providers P1, P2 and P3, respectively. The link speed between each antenna and the edge router is 20Mbps, the delay is 10ms, and Droptail queue.

A mobile (User1) walks through the avenue handing over according to a specific handover decision policy, totalling a 15 minutes walk (common speed of 4 km/h, i.e., 1.11 m/s). Two decision policies were adopted. In the first policy, the mobile prioritized signal strength, as commonly simulated in the state-of-the-art, and in the second policy the mobile prioritized the amount of credits offered by each provider, using MYHand. Provider P1 offers 30 credits for each minute that the mobile stay connected, Provider P2 offers 45 credits and Provider P3 offers 60 credits. Three different amount of credits were simulated such that the mobile could gain more or less credits according to the policy (prioritizing credits or not).

To generate traffic in the scenario, each simulation had 3 fixed users in each antenna, at most 10 meters away from the antenna, downloading a 100 Kbps constant bit rate whose source was the host Content Provider. To verify the influence of the traffic in the total of credits gained, other simulations were realised with 6, 9, 12 and 15 users connected in each antenna. In the mobile a VoIP connection, whose peer was the host VoIP peer, was simulated. The link speed between hosts Content Provider and VoIP peer, and the core network is 100Mbps, the delay is 2ms, and Droptail queue. The total of received bytes and gained credits were measured in each simulation. Varying number of users and the policy, 10 different simulations were executed.

The propagation model used was shadowing. The loss exponent and the shadowing deviation parameters were, respectively, 3.2 and 4, characterizing an external environment of an urban area, according to [18]. The MAC layer was configured to the IEEE 802.11g standard by following the parameters used in [19]. The routing protocol used was NOAH [20], suitable for the infrastructure mode.

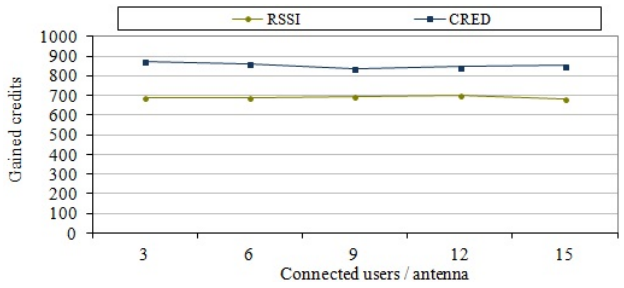


Fig. 8. Results of gained credits

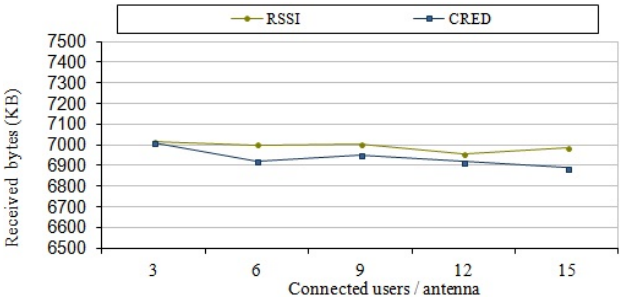


Fig. 9. Results of received bytes

The results of gained credits by the User1 as a function of the number of connected users are shown in Fig. 8. In all simulations in which the amount of credits was prioritized, the mobile User1 gained more credits than RSSI prioritization (between 20.7% and 26.7% more). The increase in the number of users, and traffic, did not affected the amount of gained credits, because it was not verified a drop or an increase trend.

Fig. 9 shows the amount of received bytes as a function of the number of connected users. In the simulations prioritizing credits, the amount of received bytes decreased compared to RSSI prioritization but it was 1.38% in the worst case, not a significant loss.

The same 10 simulations were duplicated by changing the credits offered by providers P1, P2 and P3, respectively, 10,

 ISBN: 978-1-61208-279-0

20 and 30 credits. The results were similar to the former.

The cost to the mobile for gathering incentive information was not measured because the architecture is not implemented, but it would be a few dozen of bytes. These informations will be received together to other basic schema informations, there being no need for scanning.

The handover process can be divided into three stages: decision, initiation and execution, and all these stages have dynamic features. These features do not affect MN ability to identify each network as belonging to a provider to which it has contract. The MN identifies the provider by means the SSID and roaming partners information, provided in the MIH_Get_Information.response message.

V. Conclusion and Future Work

Next Generation Networks (NGN) empower the users of mobile communication devices to opportunistically navigate through different access networks. Network selection can change according to the circumstances of offered services and required transmission parameters. An extended AAA mechanism provides on-demand connectivity to the devices even without a pre-established access plan. Specific information on the available networks and offered services must be provided to the handover decision mechanism on the mobile.

This paper presented the MYHand architecture, for providing the mobile devices with extended information for performing conscious alternative handover decisions. The architecture combines the use of the Y-Comm model and the IEEE 802.21 protocol, which is incremented with an extended schema. The MYHand architecture does not specify a decision algorithm, but assists in the decision-making process performed at the mobile device.

Validation results show an increase of 26.7% in the gained credits by using the MYHand architecture, compared to signal strength prioritization, as proposed by other works in the state-of-the-art. A decrease in the throughput using the new architecture was observed but it was less than 1.4%.

According to the MYHand architecture, different incentives and negotiation procedures can be used in the network selection mechanism, as exemplified by the rank-based model presented in this paper.

As future work, we intend to extend the architecture to include other informations, and to implement policy and access parameters negotiation between mobile and network.

Another possible future work is simulating a heterogeneous scenario with Wi-Fi and LTE antennas and evaluating the benefits of the new architecture in terms of amount of resources, robustness of RDF/OWL ontology, and the reliability if occurred changes in terms of territorial coverage area, speed of the user' moving, etc.

Acknowledgment

The authors would like to thank the National Science and Technology Institute for Critical Embedded Systems (INCT-SEC, Brazil) for funding this work by means of the following agencies: CNPq (grants # 573963/2008-8 and # 451487/2013-3) and FAPESP (grant # 2008/57870-9), and UDESC.

References

[1] ITU-T Recommendation Y.2001, "General Overview of NGN," International Telecommunication Union, Tech. Rep., 2004.

[2] J. Schiller, *Mobile Communications*, 2nd ed. Addison Wesley, 2003.

[3] J. McNair and F. Zhu, "Vertical handoffs in fourth-generation multinetwork environments," *Wireless Communications, IEEE*, vol. 11, no. 3, pp. 8–15, 2004.

[4] G. Mapp, et al., "Exploiting location and contextual information to develop a comprehensive framework for proactive handover in hetegeneous environments," *Computer Networks and Communications*, vol. 2012, pp. 1–7, February 2012.

[5] ___, "IEEE standard for local and metropolitan area networks - part 21: Media independent handover services," IEEE Std 802.21-2008, January 2009.

[6] G. Mapp, F. Shaikh, D. Cottingham, J. Crowcroft, and J. Baliosian, "Y-comm: a global architecture for heterogeneous networking," in *Proceedings of the 3rd international conference on Wireless internet*, ser. WICON '07. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2007, pp. 22:1–22:5. [Online]. Available: http://dl.acm.org/citation.cfm?id=1460047.1460075

[7] J. Wu, S. Yang, and B. Hwang, "A terminal-controlled vertical handover decision scheme in IEEE 802.21-enabled heterogeneous wireless networks," *International Journal of Communication Systems*, vol. 22, no. 7, pp. 819–834, Jul. 2009.

[8] C. Cicconetti, F. Galeassi, and R. Mambrini, "A software architecture for network-assisted handover in IEEE 802.21," *Journal of Communications*, vol. 6, no. 1, pp. 44–55, 2011.

[9] S. Rizvi, A. Aziz, and N. Saad, "An overview of vertical handoff decision policies for next generation wireless networks," in *Circuits and Systems (APCCAS), 2010 IEEE Asia Pacific Conference on*, dec. 2010, pp. 88 –91.

[10] M. Kassar, B. Kervella, and G. Pujolle, "An overview of vertical handover decision strategies in heterogeneous wireless networks," *Computer Communications*, vol. 31, pp. 2607–2620, June 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1379906.1380018

[11] F. Zhu and J. McNair, "Multiservice vertical handoff decision algorithms," *EURASIP Journal on Wireless Communications and Networking*, vol. 2006, pp. 1–13, April 2006.

[12] S. Aust, D. Proetel, N. A. Fikouras, C. Paupu, and D. Gorg, "Policy based mobile ip handoff decision (polimand) using generic link layer information," in *Proceedings of the 5th IEEE Int. Conf. Mobile and Wireless Communication Networks*, October 2003.

[13] A. Mateus and R. Marinheiro, "A media independent information service integration architecture for media independent handover," in *Ninth International Conference on Networks (ICN)*, april 2010, pp. 173 –178.

[14] A. Dutta, S. Madhani, and T. Zhang, "Network discovery mechanisms for fast-handoff," in *3rd International Conference on Broadband Communications, Networks and Systems, BROADNETS 2006*, oct. 2006, pp. 1 –11.

[15] F. Buiati, L. Villalba, D. Corujo, S. Sargento, and R. Aguiar, "IEEE 802.21 information services deployment for heterogeneous mobile environments," *Communications, IET*, vol. 5, no. 18, pp. 2721 –2729, 16 2011.

[16] R. Lopes, et al., "Exploring user's habits and virtual communities to improve ip-connectivity management," in *International Conference on Ultra Modern Telecommunications Workshops, ICUMT '09.*, oct. 2009, pp. 1 –6.

[17] G. Mapp, et al., "Exploring efficient imperative handover mechanisms for heterogeneous wireless networks," in *International Conference on Network-Based Information Systems, NBIS '09.*, aug. 2009, pp. 286 – 291.

[18] "Network Simulator - NS2," http://nsnam.isi.edu/nsnam, retrieved: April, 2013.

[19] D. F. K. Medepalli, P. Gopalakrishnan and T. Kodama, "Voice capacity of IEEE 802.11b, 802.11a and 802.11g wireless lans," in *Proceedings of IEEE Global Telecommunications Conference - Globecom 2004*, vol. 3, 2004, pp. 1549–1553.

[20] "The no ad-hoc routing agent (NOAH) website," http://icapeople.epfl.ch/widmer/uwb/ns-2/noah/, retrieved: April, 2013.

 ISBN: 978-1-61208-279-0

Reduced Complexity Decision Feedback Channel Equalizer using Series Expansion Division

S. Yassin, H. Tawfik
Department of Electronics and Communications
Cairo University Faculty of Engineering
Cairo, Egypt
syassin@eece.cu.edu.eg, Hazim.Tawfik@eece.cu.edu.eg

***Abstract*—Decision Feedback Equalizers (DFE) are used to eliminate the effect of Inter Symbol Interference (ISI) in band limited channels. The Recursive Least Squares (RLS) is one of the algorithms that are used to update the coefficients of the equalizer due to its fast convergence. However, its complexity is in the order of $O(N^2)$ multiplication per iteration. Fast fixed order algorithms are used to solve the RLS algorithm using explicit set of equations. Therefore, their complexity is in the order of $O(N)$. However, divisions are needed to perform these algorithms, making it difficult to choose the one with lower computational cost. In this paper a new metric is proposed to calculate a *weight score* for any algorithm. To verify the validity of the proposed metric, the algorithms are compared by a recent study of RLS based on Decent Coordinate Descent (DCD) iterations using computer simulation. In addition, a new method to optimize the weight score is proposed, by using multiplicative techniques to perform divisions required for the fast fixed algorithms. Results have been verified by implementing a DFE on Field Programmable Gate Array (FPGA).**

Keywords-reduced complexity; channel equalization; signal processing; division; multiplication.

I. INTRODUCTION

ISI is a common phenomenon encountered when recovering band limited channels. ISI occurs if modulation bandwidth increases beyond the channel coherence bandwidth [1]. Channel equalization is used to compensate for ISI at the receiver to decrease the bit errors. In the family of DFE equalizers the previous output decisions influence the current estimated symbol [2]. Therefore, DFE has better tracking performance than the family of linear equalizers when the channel has severe distortion and many nulls in the pass band. Both linear and non-linear equalizers are identified by a structure and an algorithm as shown in Figure 1. In the following paragraph, we will describe the structure and algorithm of DFE briefly.

The structure of DFE may be linear transversal or lattice. Both types will be treated in a similar approach to evaluate their complexity. The choice of an algorithm to update the equalizer weights is of great importance. Rate of convergence, misadjustment, computational complexity, and numerical properties are used to evaluate different algorithms [1]. The main focus of this paper is to evaluate the computational cost or complexity for DFE that is generic enough to be used for DFE with any algorihm and structure. A comprehensive survey on DFE can be found in [2].

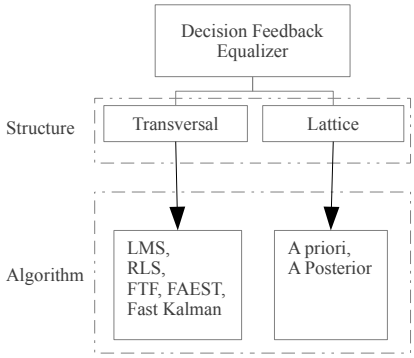


Fig. 1. DFE is classified by a structure and an algorithm

Related work such as [3], [4], and [5] are focused on the complexity reduction of DFE, because reduction in complexity leads to reduction in power consumption. This is desirable for applications with battery operated devices, where power consumption is main concern. In [3], the complexity is reduced by observing that in communication systems the input data has shift structure. In [4], spectral factorization is used to calculate the DFE coefficients in fast and efficient way. In [5], a new Dichotomous Coordinate Descent (DCD) algorithm is used to solve the RLS equations. Related work propose different solutions, but they are all have one thing in common. The common is the reduction of the total number of required mathematical operations.

The work in this paper is two fold. First, a new metric to evaluate the complexity of different DFE. This would decrease the effort to choose a suitable algorithm to be implemented. Second, DFE algorithms containing division operations consume large chip area, and it is advised to avoid them for implementation in programmable devices [6]. This work proposes a new method to implement division process. Hence, it is now permitted to use DFE containing division operations under reduced complexity constraints.

The rest of this paper is organized as follows; in Section II, candidate DFE algorithms are compared from computational cost point of view as reported in the literature. Then a new metric is proposed to evaluate the computational cost between those algorithms, and the comparison is repeated to validate the proposed metric. In Section III, it is proposed to

 ISBN: 978-1-61208-279-0

use multiplicative techniques to perform division operations involved in DFE algorithm. Therefore, the DFE weight score will be reduced, where a smaller weight score corresponds to less complexity. In Section IV, results of comparison using the proposed metric are verified using computer simulations. In addition, to verify that the weight score of division is reduced, one algorithm is implemented into FPGA using hardware description language. Resource utilization for the number of used cells in the FPGA is also reported to show the number of used cells by a DFE. Conclusion is presented in Section V.

II. Computational Cost of DFE-RLS

A. Complexity as Number of operations

In the literature of adaptive filtering such as [2] and [4], the computational cost of DFE is reported as the number of four basic mathematical operations in addition to square root. To reduce complexity it is advised neither to use divisions nor square root, and as few multiplications as possible [5]. In this paper the use of division will be permitted by reducing its weight score as will be discussed in Section III.

Let N be the total number of forward and backward taps of DFE and consider two algorithms A and B. Algorithm A has $O(7N)$ multiplications and two divisions, while algorithm B has $O(8N)$ multiplications and one division. Note that $O(7N)$ represents that the total number of an operation is dominated by a term that is seven times the number of taps in DFE. It is not straightforward to determine which algorithm has lower computational cost. Hence, it is required to find a metric or *weighting score* for each basic arithmetic operation to be able to determine the overall computational cost. The value of the weight score will represent the overall complexity of DFE. A lower weight score corresponds to reduced complexity.

In previous work where the complexity of DFE is addressed, the weight score is reported as the number of basic mathematical operations [4], [5], and [6]. The complexity of DFE is usually reported for real data and linear filtering. These results are extended to include complex data, which is the general case for communication systems [7] as shown in Table I. In [8], it was shown that the computational cost of DFE is the same as linear filtering for the same total number of taps. Hence, the calculations are not only applied for linear filtering but also for the DFE problem.

Candidate algorithms will be chosen to represent the families of conventional Least Mean Square (LMS), conventional RLS, fast fixed order, lattice, and the family of DCD. Each candidate algorithm is generic enough to represent its family. The same procedure can be used to calculate the weight score for any other algorithm. Due to the desirable feature of fast fixed order family [3] with the complexity in the order of $O(N)$, three members of this family will be evaluated; namely Fast Transversal Filter (FTF), Fast Aposterior Error Sequential Technique (FAEST) and Fast Kalman. In the rest of this subsection, each family of algorithms will be discussed briefly.

(LMS) algorithm has the lowest computational cost [2]. Although LMS has the slowest convergence rate, it will be used as a reference scenario for the proposed metric.

TABLE I
Computational Cost of DFE using RLS Algorithm for general complex data where N is the total number of taps

Algorithm	$\times$	+	$\div$
Conventional	$4N^2+16N+1$	$4N^2+12N-1$	1
Apriori Lattice	$64N$	$32N$	$32N$
FAEST	$28N+6$	$28N+2$	5
FTF	$28N+10$	$28N+1$	3
Fast Kalman	$36N+2$	$32N+1$	2
ERLS-DCD-16	$12N$	$134N$	0
LMS	$8N+2$	$8N$	0

Fast fixed order algorithms are based on RLS; namely FTF, FAEST and Fast Kalman. Their complexity is in the order of $O(N)$ and their rate rate of convergence is similar to the conventional RLS, which is considered fast. Therefore, they will be suitable for systems needing short iterations to reach optimum weights of the channel to reduce transmission overhead [1].

The lattice RLS algorithms have a lot of desirable features such as improved numerical properties and modularity [6]. However, lattice filters have higher computational requirements and can not be used in all applications [5]. They will be compared here with other families, to show the effectiveness of using the proposed metric. It will be shown in Section IV how the proposed metric simplifies the comparison between different families.

A recent study to reduce the complexity of RLS algorithm is based on DCD iterations [5]. The RLS is expressed in terms of auxiliary normal equations with respect to increments of the filter weights. Auxiliary equations are solved using line search methods. These methods have more than one solution for conventional RLS problem. One of these solutions is chosen, which has the least computational cost among its family namely; Exponentially Weighted RLS with 16 iterations per sample (ERLS-DCD-16). However, the ERLS-DCD offer reduced complexity at the expense of low convergence time.

After describing the candidate algorithms briefly, it can be observed from Table I that it is not straightforward to determine which algorithm has the least computational cost. Moreover, it is impossible to arrange the rows in Table I in a descending order according to computational cost. Therefore, we found a need to develop a new metric to evaluate the weight score of DFE. To accomplish this goal, it is proposed to normalize all operations to a weight score as will be discussed in Subsection II-B.

B. Complexity Weight Score

As stated in [9] the simplest mathematical operation is the binary addition. Therefore, all operations should be mapped to a finite number of additions. Then, the overall weight score of DFE will be a linear summation of the weight score of all used operations. In order to accomplish the normalization,

 ISBN: 978-1-61208-279-0

TABLE II
MULTIPLICATION USING ADD AND SHIFT METHOD

Multiplicand	$\ldots\ldots\ldots x_3\ x_2\ x_1\ x_0$
Multiplier	$\ldots\ldots\ldots y_3\ y_2\ y_1\ y_0$
Partial Products	$\ldots x_3y_0\ x_2y_0\ x_1y_0\ x_0y_0$
	$\ldots x_3y_1\ x_2y_1\ x_1y_1\ x_0y_1$
	$\ldots x_3y_2\ x_2y_2\ x_1y_2\ x_0y_2$
	$\ldots\ldots\ldots\ldots\ldots\ldots$
Final Product	$Z_{M_c+M_l-1}\ldots\ldots\ldots\ldots\ldots\ Z_2\quad Z_1\quad Z_0$

we propose to analyse the low level details of each operation involved in DFE. In the following, the normalization will be developed for the multiplication operation. This normalization enables one to compare fairly between different families of algorithms.

As described in Table II the binary multiplication may be performed by adding the multiplier to the multiplicand and storing the result. Then the multiplier is shifted one bit to the left and added to the previous result [9]. To illustrate the multiplication of X and Y Let M_l and M_c be the data width for multiplier and multiplicand respectively. The first partial product is the binary multiplication of first bit in the multiplier y_0 and each bit in the multiplicand x_i where i is in the range 0 to M_{c-1}. Recalling that binary multiplication is the logical *"AND"* operation, the first bit of the first partial product equals x_0y_0, and the second bit of the first partial product equals x_1y_0, up to $x_{M_c-1}y_0$. The same procedure is repeated for each bit in the multiplier until all of the partial products are generated.

This method is the simplest from complexity point of view because it can be implemented using only one adder and one shifter. However, from processing time point of view it is considered the slowest to store the final result.

Another multiplication method is the Wallace Tree [9]. It is based on parallel generation of the required number of partial products. Afterwards this number of partial products is reduced. It is the fastest multiplier scheme at the expense of increased computational cost. A method that is considered a good compromise between processing time and computational cost is the iterative array of cells [9]. Hence, the iterative array multiplication will be used in the rest of this paper. However, the new proposed metric can be calculated for all multiplication methods as will be discussed in the following paragraph. In general, the iterative array method is used for short data lengths, which is the case for communication receiver [10], because its delay increases with operand length [9].

Multiplication consists of a finite number of cells or building blocks. To perform one multiplication, a finite number of building blocks are needed. let G be the number of building blocks needed to perform one multiplication. In order to calculate the number of required building blocks G, the following

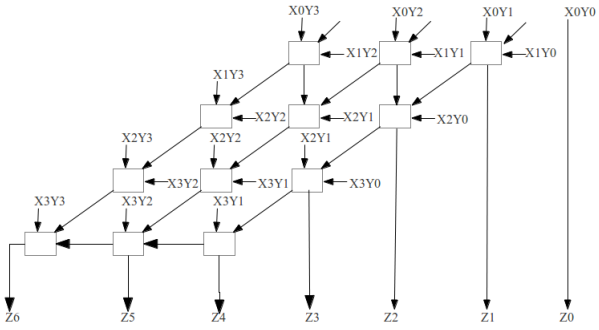


Fig. 2. Structure of 4 x 4 multiplier using full adder cells

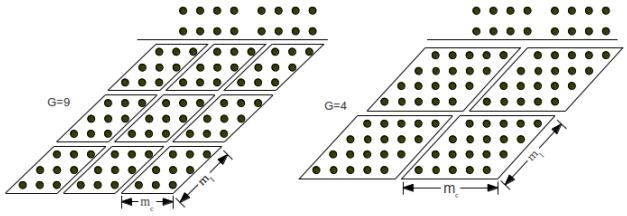


Fig. 3. Compromise between G and $m_l \times m_c$. Each black circle is used to represent one adder cell.

relation is used [9]

$$G = \lceil \frac{M_l \times M_c}{m_l \times m_c} \rceil \quad (1)$$

Where and m_l and m_c are the data width of the building block used. For example, using $M_l = M_c = 12$ and $m_l = m_c = 4$ will result in $G = 9$. Therefore, nine building blocks will be needed to construct one multiplier and each building block is $m_c \times m_l$ multiplier.

To complete the normalization, each building block is composed of finite number of 1-bit full adder cell as shown in Figure 2. Hence, one multiplier building block is composed of $[(m_l - 1) \times (m_c) + 1]$ 1-bit full adder blocks. Therefore, the weight cost of multiplication M_{score} can be rewritten as

$$M_{score} = ((m_l - 1) \times m_c + 1) \times G \quad (2)$$

Equation 2 explains how to calculate the weight score of multiplication operation in terms of 1-bit full adder cells. In Section III, the weight score of division operation will be calculated in terms of adder cells similarly.

It can be observed that there exist an inverse proportional relation between the data width of the building block and the number of building blocks G. As $m_l \times m_c$ increases, the number of required building blocks G decreases as shown in Figure 3. From complexity point of view, the weight score of multiplier is independent of the number of used building blocks G. Intuitively it does not matter whether to use one large building block or many small building blocks. In both situations the total number of 1-bit full adder cells are the same as shown in Figure 3. However, from processing time point of view it is favoured to use smaller building blocks. This is due to the fact that the processing time to have a result is proportional to the width of the building block [11].

 ISBN: 978-1-61208-279-0

It can be observed, from Figure 3, that using 3×3 building block generates an extra partial product because the data width is not integer multiple of the building block data width. This extra partial product will be zeros and may be skipped using a multiplexer [11]. In general, for any $M_l \times M_c$ data width, and any $m_l \times m_c$ building block width the extra partial products will not affect the final weight score. This is because the weight score is used for comparison based on the number of used 1-bit full adder cells and not on the number of building blocks.

III. Division using series expansion

Division algorithms can be divided into two categories; the first category is based on iteration of subtraction and the second is based on iteration of multiplication [9]. The first category is performed as the normal pencil and paper division. At each iteration, there is remainder R, divisor D, and quotient Q. The ith bit of quotient q_i can be calculated using the following equation

$$R(i) = R(i+1) - q_i * D * 10^i \tag{3}$$

For example, the first iteration of division of 4000 over 3 will be $4000 - 1 \times 3 \times 1000$ where $R(i+1) = 4000$, $q_3 = 1$, $D = 3$ and $10^i = 10^3$. Each iteration q_i is chosen to be 0 or 1 according to the negative or positive value of $R(i)$ respectively. This method is considered slow because the delay is proportional to the ratio between the divisor and the dividend. Therefore, we will consider another category of division algorithms, which has less processing time.

The second category of division algorithms is based on the use of Maclaurin series expansion [9]. The division $\frac{a}{b}$ will be obtained as the multiplication of a and $\frac{1}{b}$. Note that, according to the floating point standard [12], numbers are represented in either 32 or 64 bits with the format $1.x \times radix^{exponent}$ where x is a fraction. Therefore, one can use $b = 1 + x$ and ignore higher orders of x, depending on the required accuracy, in the familiar series expansion

$$\frac{1}{b} = \frac{1}{1+x} = \underbrace{(1-x)(1+x^2)(1+x^4)(1+x^8)}_{Memory location Address} \tag{4}$$

All possible values of the reciprocal are stored in a memory element with its length proportional to the data width. This would replace the need to true division with a simple memory allocation. In typical communication systems [10] the data widths are in the range 8, 10, 12 or 14 bits, because a memory block of size 2^c is needed to store all values of a digital word with data width c bits. Hence, the memory size needed is approximately 0.25, 1, 4 or 16 kbits.

To form one divider a finite number of adders and multipliers is needed to form the memory address. This number varies according to the desired data accuracy. For 8-bit data accuracy, six multipliers and four adders are needed to implement one divider and the calculation of the memory address is shown in Figure 4. To calculate the memory element address we need to calculate the powers x^2, x^4, and x^8 depending on the desired

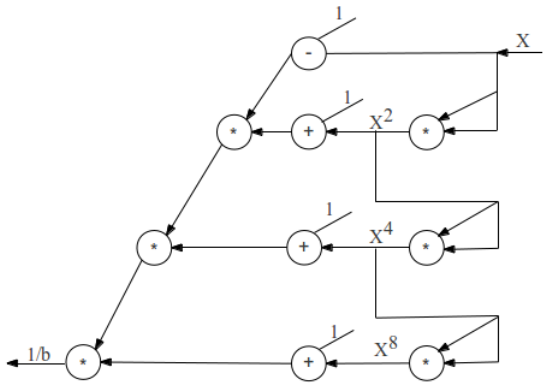


Fig. 4. Division using series expansion for 8-bit accuracy.

accuracy. For 8-bit accuracy, the weight score of one division D_{score} can be calculated according to the following relation

$$D_{score} = 6 * M_{score} + 4 \tag{5}$$

To achieve higher accuracy for the division operation, higher powers of x should be considered. For increased accuracy, such as 12-bit, x^{12} can be calculated by multiplying x^4 by x^8, which have been calculated earlier as shown in Figure 4. Consequently, one extra multiplier and two adders are needed. The weight score can be modified accordingly to be

$$D_{score} = 7 * M_{score} + 6 \tag{6}$$

According to equations 5 and 6 the division is normalized into a finite number of 1-bit full adder cells. In the following section results of weight score for different algorithms are presented.

IV. Simulation Results

In this section, results obtained by computer simulation are presented. The complexity weight score of different algorithms is plotted against filter order N. First, the weight score of each operation is calculated according to equations 2 and 5. For each algorithm the total computational cost is calculated according to Table II. Then the number of multiplications and divisions in the table is multiplied by M_{score} and D_{score} respectively. In this manner the computational cost is compared fairly. The total weight score of each algorithm is the linear summation between the weight score of additions, multiplications and divisions included in that algorithm. This procedure is repeated for different values of equalizer order N as shown in Figure 5.

The conventional RLS was not plotted due to its high weight score that would compress all other graphs at the bottom of the y-axis. The weight score of conventional RLS is almost $5.5k$ at $N = 10$. However, its complexity increases exponentially afterwards until it reaches $20k$ at $N = 20$.

It is observed that the LMS has the least weight score over all values of N. This result was expected as was mentioned in Section II-A. The ERLS-DCD-16 has slightly higher complexity than LMS. This result conforms to the results in [5],

 ISBN: 978-1-61208-279-0

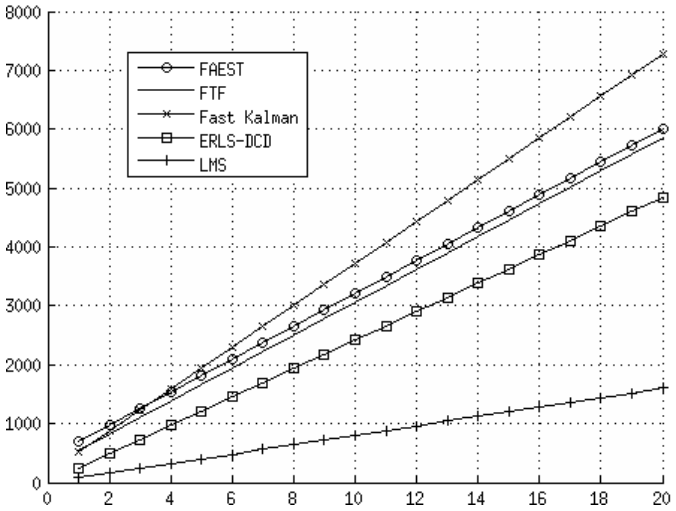


Fig. 5. Complexity of DFE versus the total number of taps

TABLE III
FPGA RESOURCES FOR THE PROPOSED ENHANCEMNT FOR FTF

Resources	$N = 8$	$N = 12$	$N = 16$
Slice	127(1%)	177(1%)	206(1%)
D-FF	136(1%)	186(1%)	226(1%)
LUT-4	206(1%)	313(1%)	403(1%)
BRAM	1(1%)	2(2%)	4(4%)
DSP48	2(2%)	3(3%)	3(3%)

which states that LMS complexity is proportional to $2N$, while ERLS-DCD-16 complexity is proportional to $3N$. Both FTF and FAEST have almost the same weight score over N. It is important to observe that fast Kalman has higher weight score than FASET. This result was not obvious before using the weight score metric because FAEST has 5 divisons while Fast Kalman has 2 divisions only.

In order to verify the usage of the proposed division method experimentally, the resource utilization for the Xilinx FPGA "Spartan3A-DSP1800" [13] is shown in Table III. Results are obtained using Xilinx development suite ISE12.1. Using the automatic synthesis procedure, the synthesizer fails to design the division process and ends with error messages. This result was expected because the implementation of division operation into FPGA is problematic [14]. Different division implementations such as restoring, non restoring, and Xilinx Intellectual Property(IP) consumes large area of the FPGA chip. Afterwards, divisions are implemented as proposed using memory elements, which is called Block RAM (BRAM). The synthesis process succeeded with low chip utilization . Note that the multiplication and the tapped delay line of the DFE are implemented using the DSP blocks of the FPGA [13].

V. CONCLUSION

DFE equalizers are used to eliminate ISI phenomena. Many algorithms are used to adopt the coefficients of equalizers such as conventional RLS. However, RLS complexity is high and in the order of $O(N^2)$. Other algorithms have lower complexity such as fast RLS algorithms with computational cost in the order of $O(N)$, and ERLS-DCD-16. However, it is not straightforward to compare fairly between different algorithms. In this paper a new metric, *weight score*, is proposed to measure the complexity of each algorithm. Each algorithm is normalized as a function of the number of 1-bit full adder units needed to construct the DFE. Comparison results were presented using computer simulation. In addition, a new method was proposed to decrease the weight score of the division operation. This is accomplished by using the iterative array division, which is based on Maclaurin series expansion. To verify the reduction of weight score experimentally one algorithm, (FTF), is implemented in FPGA and resource utilization results were presented. It was shown that division operation can be implemented into FPGA with low chip area utilization.

ACKNOWLEDGMENT

This work is supported by the Egyptian National Telecommunication Regulatory Authority (NTRA).

REFERENCES

[1] T. Rappaport, Wireless Communication: Priciples and Practices, 2nd ed., New Jersey: Prentice Hall, 2001, pp. 308-318
[2] J. Proakis, "Adaptive Equalization for TDMA Mobile Radio", IEEE Transactions on Vehicular Technology, vol. 40, no. 2, May. 1991, pp. 333-341
[3] J. Cioffi and T. Kailath, "Fast, Recursive Least-Squares Transversal Filters for Adaptive Filtering", IEEE Transactions on Acoustics, Speech, and Signal processing, vol. 32, Apr. 1984, pp. 304-337
[4] N. Al-Dhahir, J. Cioffi, "Fast computation for Channel-Estimate based Equalizers in Packet Data Transmission", IEEE Transactions on Signal Processing, vol. 43, no. 11, Nov. 1995, pp. 2462-2473
[5] Y. Zakharov, G. White, and J. Liu, "Low-Complexity RLS Algorithms Using Dichotomous Coordinate Descent Iterations", IEEE Transactions on Signal Processing, vol. 56, no. 7, July. 2008, pp. 3150-3161
[6] H. Sayed, Fundamentals of Adaptive Filtering, 2nd ed., New Jersey: Wiley-IEEE Press, 2003, pp. 600-620
[7] B. Bjerke, J. Proakis, K. Martin Lee, and Z. Zvonar, "A Comparison of GSM Receivers for Fading Multipath Channels with Adjacent- and Co-Channel Interference", IEEE Journal on selected areas in communications, vol. 18, Nov. 2000, pp. 2211-2219
[8] Y. Yang, X. Gao, Z. Gao, and X. Wang, "An Classification-based Adaptive Decision Feedback Equalizer for Rayleigh Multipath Channel", Journal of Computational Information Systems, vol. 8, no. 2, Jan. 2012, pp. 869-876
[9] M.J. Flynn, and S.F. Oberman, Advanced Computer Arithmetic Design", 2nd ed., New York: John Wiley and Sons, 2001, pp. 113-125
[10] O. Dabeer, and U. Madhow "Channel Estimation with Low-Precision Analog-to-Digital Conversion", IEEE International Conference on Communications (ICC 10), vol. 2, May. 2010, pp. 23-27
[11] R. Singh, P. Kumar, and B. Singh, "Performance Analysis of 32-Bit Array Multiplier with a Carry Save Adder and with a Carry Look-Ahead Adder", International Journal of Recent Trends in Engineering, vol. 2, no. 6, Nov. 2009, pp. 83-86
[12] IEEE Standard for Floating-Point Arithmetic,"IEEE Std 754-2008", pp.2-7, Aug. 2008, doi: 10.1109/IEEESTD.2008.4610935.
[13] Xilinx Inc., "XtremeDSP DSP48A for Spartan-3A DSP FPGAs User Guide", UG431 (v1.3), July. 2008, pp.32-34
[14] N. Sorokin, "Implementation of high-speed fixed-point dividers on FPGA", Journal of Computer Science and Technology, vol. 6 no. 1, May. 2006, pp. 8-11

 ISBN: 978-1-61208-279-0

Double Directional Channel Characterization on Board Ships

H. Kdouh, H. Farhat, C. Brousseau, G. Zaharia,
G. Grunfelder, G. El Zein
Institut d'Electronique et de Télécommunications de Rennes
Rennes, France
hussein.kdouh@insa-rennes.fr

T. Tenoux, Y. Lostanlen
Siradel
Rennes, France
ttenoux@siradel.com

***Abstract*— Due to the metallic structure of decks, bulkheads and watertight (WT) doors, wireless communications are a serious challenge in the particular environment of ships. In order to deploy reliable shipboard wireless networks, wireless devices (access points, routers, sensor nodes, etc.) must be located at strategic locations ensuring full radio coverage and network connectivity. Strategic locations can be determined from the identification of the main propagation directions of electromagnetic (EM) waves within a ship. This paper presents the results of a radio propagation measurement campaign performed on board a ship. A dual-band Multiple-Input Multiple-Output (MIMO) channel sounder and antenna arrays have been used. Measurement data have been processed with a classic beamforming technique and a high resolution algorithm to extract dominant paths. A ray-tracing based simulation tool has been used to understand measurement results. Obtained results are used for optimal placement of radio devices when deploying shipboard wireless networks.**

Keywords-propagation; channel sounding; ray-tracing; ships

I. INTRODUCTION

Ships are an important part of modern systems, which are widely used in commercial and military purposes. Modern ships are equipped with automatic alarm and monitoring systems, which control and ensure safety and accuracy of the whole ship operation. Current shipboard monitoring systems use extensive lengths of cables to connect several thousands of sensors to control units. In addition to the high cost and weight due to wires installation during ships construction, vessels represent a complex and harsh environment in which extensive lengths of wires are vulnerable to heat, moisture and toxic agents [1]. Hence, applying wireless technologies such as Wireless Sensor Networks (WSN) to shipboard monitoring systems may be a cost-effective and survivable solution. Furthermore, wireless systems are easily and inexpensively reconfigured. Moreover, wireless communications can be used for other applications on board ships. Ferry companies, for example, aim to enhance their passengers' satisfaction by equipping their boats with WiFi networks. Cordless phones are very useful for communication between crew members.

However, wireless communications are a serious challenge in the particular harsh environment of ships. In fact, several factors may limit the performance of wireless systems on board ships [2]. Firstly, bulkheads and doors are made of metal, most often steel. Although the steel is not a perfect conductor, it can severely decrease the power of radio waves. Another limiting factor is the multipath propagation: a transmitted EM wave could be reflected, scattered or diffracted by different objects leading to several delayed copies to the receiver. The existence of multiple copies of the transmitted wave may cause disturbing interference signals at the receiver. In order to deploy reliable wireless networks on board a ship, wireless devices must be located at strategic locations ensuring optimal radio coverage and network connectivity. Strategic locations can be determined by identifying the main directions of EM waves within a ship.

Few works have studied the EM waves propagation on board ships [2-5]. However, spatio-temporal characterization of EM waves propagation within this particular environment has not been done yet. Existing studies have only considered the possibility of wireless communications without determining the directions of EM waves propagation. In [6], we have conducted Continuous Wave (CW) measurements on board several ferries to verify the possibility of intra-, inter-compartments and inter-decks radio communication. Measurement results helped us to determine the path loss exponents in different shipboard environments and different communication configurations. However, CW measurements cannot determine the Directions of Departure (DoD) and Directions of Arrivals (DoA) of EM waves. Double directional or MIMO measurements may be an efficient technique to estimate DoD and DoA by using antenna arrays at both link ends. In this paper, we present the results of a measurement campaign using a dual-band MIMO channel sounder [7]. This measurement system, developed in the IETR laboratory, was already used for outdoor, indoor and outdoor-to-indoor MIMO channel sounding [8]. The performed measurements provide a spatio-temporal characterization of the EM wave propagation within a ship. Collected data are analyzed using a classic beamforming technique and a high resolution algorithm. Obtained results are then compared to simulation results based on a ray-tracing tool. This comparison helps us to understand directions of propagation of EM waves in typical shipboard environments such as engine rooms, parking and passenger decks. Obtained results are used to ensure an optimal placement of radio devices when deploying shipboard wireless networks.

The remainder of this paper is organized as follows: Section II describes the methodology and measurement setup used in this study. Section III presents and analyzes the obtained measurement and simulation results. Finally, several conclusions are drawn in Section IV.

 ISBN: 978-1-61208-279-0

II. METHODOLOGY AND MEASUREMENT SETUP

The dual frequency band (2.2 and 3.5 GHz) used channel sounder transmits a spread spectrum waveform using a periodic M-sequence. It has an 11.9 ns temporal resolution for 100 MHz sounding bandwidth. The dynamic range is 50 dB for the 1023 code length. Synchronization between the transmitter (Tx) and the receiver (Rx) is achieved with highly stable 10 MHz rubidium oscillators. Different impulse response lengths can be chosen from 1.27 to 81.84 µs, depending on the sounding bandwidth and the code length. As an example, for 100 MHz bandwidth and 1023 code length, the recorded impulse response length is 10.23 µs.

Different types of antenna arrays were developed for this sounder. At 2.2 GHz, a 4-element Uniform Linear antenna Array (ULA) and a 16-element Uniform Rectangular antenna Array (URA) are used respectively for the Tx and the Rx to characterize the double directional channel on a 120° beamwidth in the horizontal plan. At 3.5 GHz, a 4-element Uniform Circular Array (UCA) is used at Tx and a 16-element UCA is used at Rx. With this configuration, we can characterize 360° azimuthal double directional channel at both link sides. In order to improve the measurement dynamic range, power amplifiers have been directly integrated in the Tx antenna array, and low noise amplifiers have been placed directly behind the Rx antenna array.

We assume a quasi time-invariant channel during the measurements. Attention was paid that no people were moving in the surrounding area. For each Rx location, several measurements were taken and averaged to reduce the noise effect. The collected channel data were stored on a laptop for post-processing.

The measurement objective was to characterize the double directional channel impulse response h [9]. In the case of omnidirectional antennas at Tx location r_{Tx} and Rx location r_{Rx}, it could be expressed as:

$$h(r_{Tx}, r_{Rx}, \tau, \theta_{DoD}, \theta_{DoA}, \phi_{DoD}, \phi_{DoA}) = \sum_{s=1}^{S} h_s(r_{Tx}, r_{Rx}, \tau, \theta_{DoD}, \theta_{DoA}, \phi_{DoD}, \phi_{DoA}) \quad (1)$$

where S is the number of multipath components, τ is the delay, Θ and Φ are the azimuth and elevation angles of DoD and DoA. With the plane wave assumption, the contribution of each multipath component s is:

$$h_s(r_{Tx}, r_{Rx}, \tau, \theta_{DoD}, \theta_{DoA}, \phi_{DoD}, \phi_{DoA}) = |a_s| e^{j\varphi_s} . \delta(\tau - \tau_s) . \delta(\theta_{DoD} - \theta_{DoDs}) . \delta(\theta_{DoA} - \theta_{DoAs}) \times \delta(\phi_{DoD} - \phi_{DoDs}) . \delta(\phi_{DoA} - \phi_{DoAs}) \quad (2)$$

where $|a_s| e^{j\varphi_s}$ is the multipath component complex amplitude of the component s.

Analysis of collected data is performed with a classic beamforming technique [10] and the high resolution Space-Alternating Generalized Expectation-maximization (SAGE) algorithm [11], in order to determine dominant paths which propagate between Tx and Rx. We gave a special attention to the DoD and DoA, which are the most important parameters to determine the directions of propagations. In order to precisely determine these directions, a simulator of EM propagation [12] has been used. This simulator, based on a 3D ray tracing algorithm [13], computes successive EM interactions thanks to the Geometrical Optics (GO) or Uniform geometrical Theory of Diffraction (UTD). This computation is associated with a 3D geometrical description of the considered scene, including the structures and the positions of bulkheads, floors, ceilings and WT doors. A maximal number of 4 reflections and one diffraction are considered for each simulation scenario. When diffraction is involved, only one reflection can occur before and after it.

III. MEASUREMENT RESULTS AND ANALYSIS

This section presents and analyzes the obtained measurement and simulation results. Three typical environments have been studied on board the "Armorique" ferry: the engine rooms, the parking and the passenger deck. The engine rooms of "Armorique" include metallic engines, generators, valves and pumps arranged in a complex way. The common bulkheads of these rooms are metallic and include sliding metallic WT doors. The parking is a big hall with metallic walls. A metallic wall located in the middle of the parking divides it into two main parts. Vehicles of different types were parked in the parking (cars, buses, trucks, etc.) when performing measurements. The passenger deck is composed of passengers' cabins and corridors. In contrast with the two previous environments, corridor walls and cabin doors are not fully made of metals.

A. EM waves propagation within engine rooms

In spite of the fully metallic structure of bulkheads and WT doors in the lower decks areas of ferries, CW measurements have shown that wireless communications between adjacent rooms remain possible after closing WT doors [6]. A MIMO channel sounder and antenna arrays are used to identify the openings allowing EM waves leakage. The studied environment is the second deck of "Armorique", where sliding metallic WT doors are used between adjacent rooms. This environment is highly metallic and confined. It generates several propagation phenomena (reflection, diffraction, scattering). Measurements were carried out using the URA and ULA arrays at 2.2 GHz. The study was limited to a 120° characterization in the azimuthal plan. It is sufficient to characterize the propagation phenomena through the bulkheads doors between adjacent rooms. Fig. 1 shows the layout of the rooms, locations of Tx and Rx in the second deck of "Armorique".

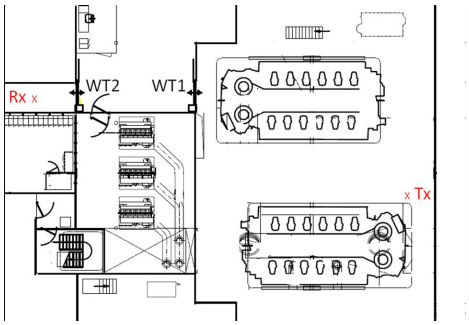


Figure 1. Tx and Rx locations in the engine rooms.

Two WT doors, called WT1 and WT2 on Fig. 1, are located on the propagation path between Tx and Rx, which are in No Line of Sight (NLoS) configuration. Measurements were carried out when WT1 was opened and when it was closed respectively, while WT2 was maintained closed. Measurement data were processed using the conventional beamforming. Fig. 2 and Fig. 3 present the normalized space-delay power graph for the two measured scenarios (radius scale of beamformer graph is in ns, and the color represents the normalized amplitude at each angle). These graphs represent the spatio-temporal channel response and allow identifying the main directions of energy propagation between Tx and Rx. It can be seen that the received energy is not homogeneously distributed on the 120° beamwidth. For the two scenarios, the angular distributions of energy are similar. It can be seen that a main beam of energy with about 20° beamwidth (between 125° and 145°) and other beams around 180° direction are formed at the Tx side. Moreover, a beam of 50° is formed around 0° at the Rx side. The main beams formed at the Tx and Rx sides correspond to the energy going to WT1 and coming from WT2 respectively. This similarity between the two scenarios (closed and open door) shows that the energy is propagating through the WT doors regardless their status. The other beams seen at 180° at the Tx side may be probably due to two respective reflections of EM waves on the bulkhead in front and behind the Tx, before penetration through WT1 and WT2 towards Rx. We have presented in [14] similar measurements for scenarios where the direct path between Tx and Rx is blocked by the WT door. Its results show that radio signals propagation is made through the openings on the edges of the metallic watertight doors.

B. *EM waves propagation within the parking*

As stated before, the parking of "Armorique" is a big hall where all walls, ceiling and floor are totally metallic. A big bulkhead installed in the middle divides the parking into two parts (called lower part and upper part in the following). Two communication scenarios were considered. In the first one, Tx and Rx are both located in the lower part of the parking (Line of Sight LoS configuration). In the second scenario, Tx is located in the upper part of the parking and the Rx is located in the lower one. The middle bulkhead blocks the LoS between the Tx and Rx (NLoS configuration).

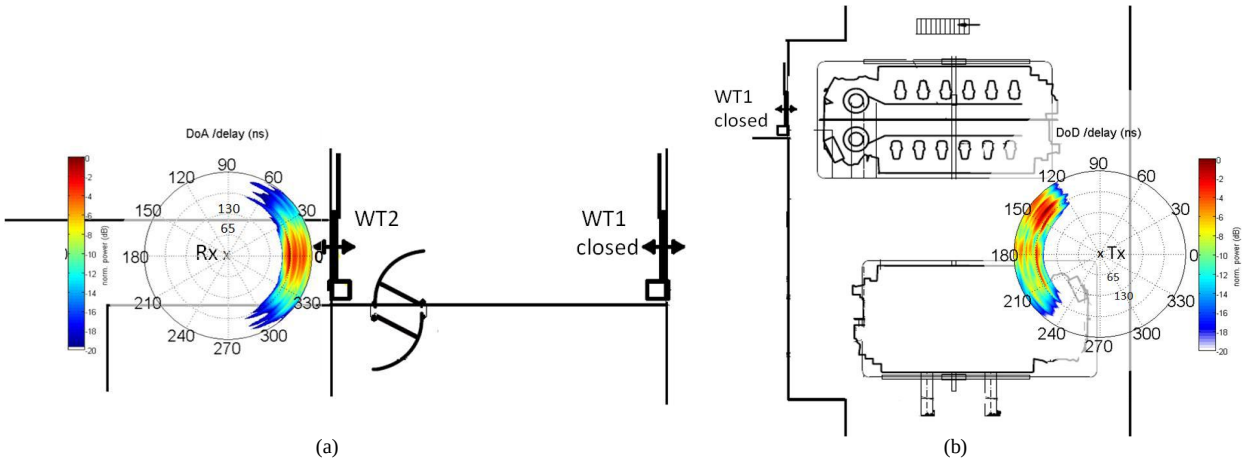


Figure 2. Beamforming results when WT1 and WT2 are closed: (a) At Rx (DoA), (b) At Tx (DoD).

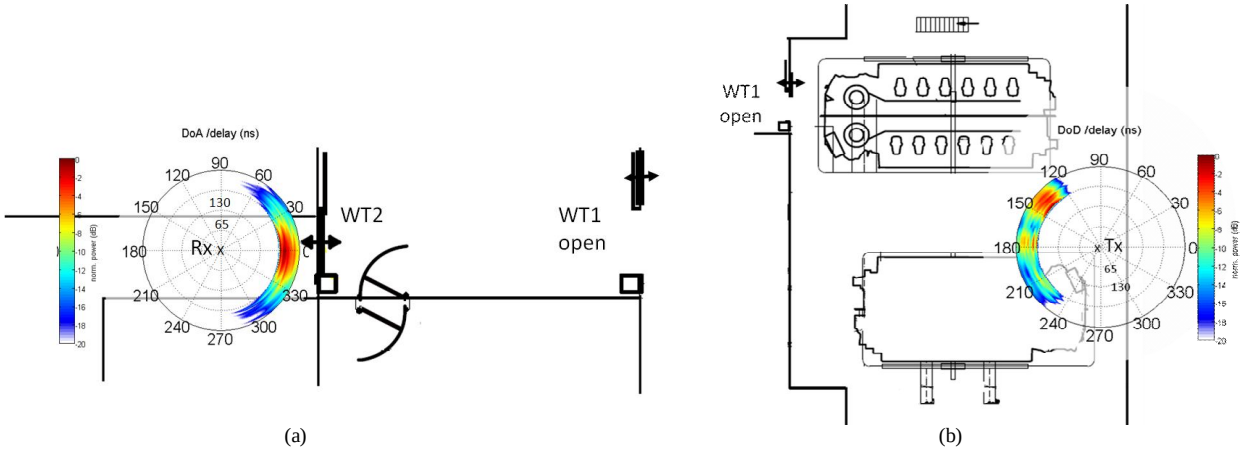


Figure 3. Beamforming results when WT1 is opened and WT2 is closed: (a) At Rx (DoA), (b) At Tx (DoD).

 ISBN: 978-1-61208-279-0

The ULA and URA 2.2 GHz are used in these measurements. The results are processed using beamforming and compared to ray-tracing simulation results.

Fig. 4 shows the layout of the parking, locations of the Tx and Rx and ray tracing results for the LoS configuration. A guiding effect is clearly observed in the lower part of the parking due to the middle bulkhead; some other paths coming from the upper part can be detected. The middle bulkhead and the wall of the parking act as a waveguide between the Tx and Rx. Fig. 5 shows the comparison between beamforming results and ray-tracing simulation results at Tx and Rx. The comparison indicates also that main beams (in red) are formed in the lower part of the parking. A significant agreement between measurement and simulation results is found.

Fig. 6 shows the layout of the parking, locations of the Tx and Rx and beamforming and ray-tracing results for the NLoS configuration. In spite of the middle bulkhead, which blocks the direct path between Tx and Rx, EM waves are able to achieve the Rx through reflections on the parking walls and the middle bulkhead, and diffraction on the middle bulkhead edges. As in the LoS configuration, a significant agreement between measurement and simulation results is found. This agreement proves that the model of the parking used for simulation can be considered as realistic.

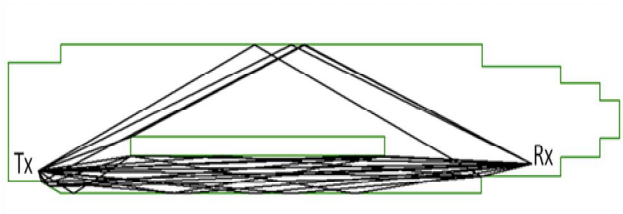
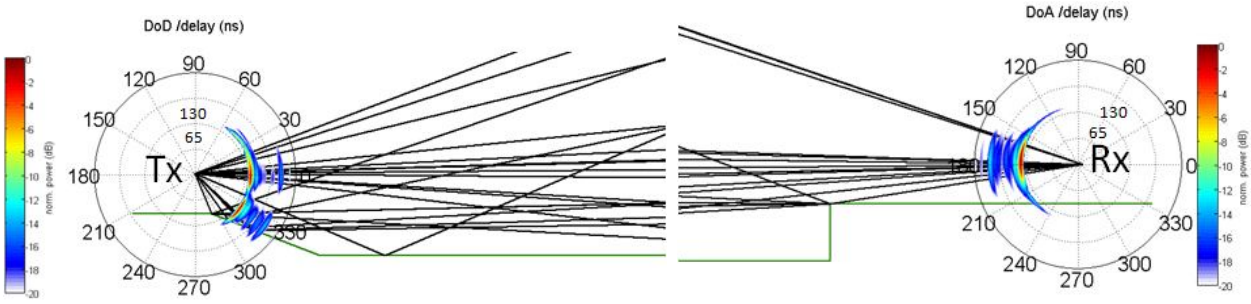


Figure 4. Ray-tracing results for the parking in LoS configuration.

C. EM waves propagation on the passenger deck

Fig. 7 shows the layout of the passenger deck and the locations of Tx and Rx. Rx is placed in the stairway located at the top left of this deck. The stairway door is opened. Tx is placed at three locations (Tx1, Tx2 and Tx3 in Fig.7).Tx is firstly located in a passenger cabin whose door is closed (location Tx1), then in a corridor in the middle of the deck (location Tx2) and finally at the bottom right corner of the deck (location Tx3). In all these measurement scenarios, Tx and Rx are in NLoS configuration. As this environment is not totally metallic like the parking and the engine rooms, no particular propagation directions can be excluded. Thus, the circular arrays have been used at the Tx and Rx to determine the directions of propagation in the 360° azimuth.



(a) (b)

Figure 5. Beamforming and ray-tracing simulation results for LoS configuration in the parking: (a) at Tx (DoD), (b) at Rx (DoA).

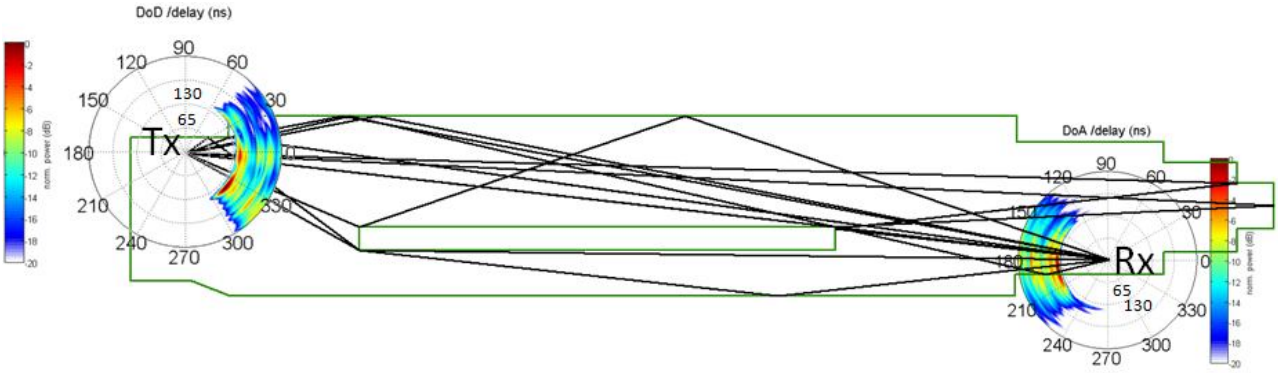


Figure 6. Beamforming and ray tracing results for the parking in NLoS configuration.

 ISBN: 978-1-61208-279-0

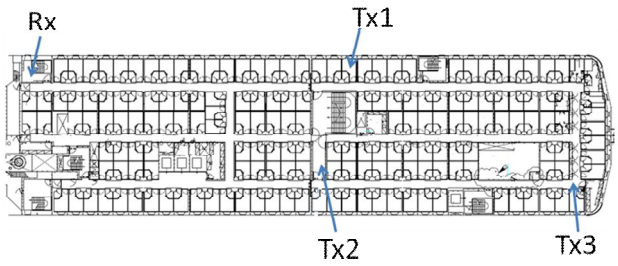


Figure 7. Layout of the passenger deck of "Armorique" with the locations of transmitter (Tx1, Tx2 and Tx3) and the receiver (Rx).

The propagation paths parameters are extracted using SAGE algorithm. Fig. 8 presents the normalized DoD and DoA results for all measurement positions. Black arrows indicate directions and the length of each one indicates its normalized amplitude (one tick interval corresponds to 10 dB). This overview shows that for the DoD, some privileged propagation directions can be easily identified. Concerning the DoA, for the three Tx locations, we notice that received energy directions are more equally distributed on 360°. This can be explained by the receiver location in the stairway. Due to its metallic walls, this environment is similar to a reverberant chamber. One can think that EM waves will arrive mainly through the entrance door of the stairway. After that, EM waves are reflected on the metallic walls of stairway and achieve the receiver from all 360° directions.

An adaptation of the ray-tracing based simulation tool was made at 3.5 GHz, mainly concerning the dielectric properties of materials constituting the passenger deck (used for the computation of the reflection, transmission and diffraction coefficients).

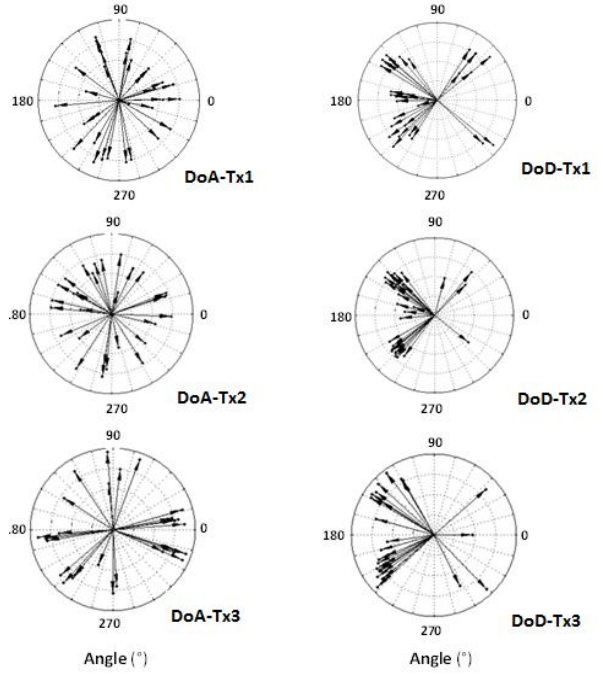


Figure 8. DoA and DoD for the three measurement scenarios on the passenger deck.

For the ray-tracing simulations, by assuming that the propagation is made through the corridors, and because of the current limitation of the RT simulator to compute complete paths from Tx to Rx, a virtual point of reception is located in the corridor near the Tx position. The idea is to visualize how the radio waves reach this point from the Tx. Fig. 9 presents the comparison between the DoD obtained from the measurements (black arrows) and ray-tracing tool (blue lines) for the Tx1 location. In this measurement scenario, Tx was located in a passenger cabin. The main directions of energy propagating from the passenger cabin to Rx position obtained from ray-tracing present a significant agreement with the measurement results. The obtained results validate our assumption that the propagation in this environment mainly occurs through the corridors. These directions show that EM waves penetrate through the thin walls of cabins (which are not metallic) before arriving to the corridor, and then they are guided to the Rx (through reflections). Note that the wall located at the left of the figure corresponds to a metallic structure.

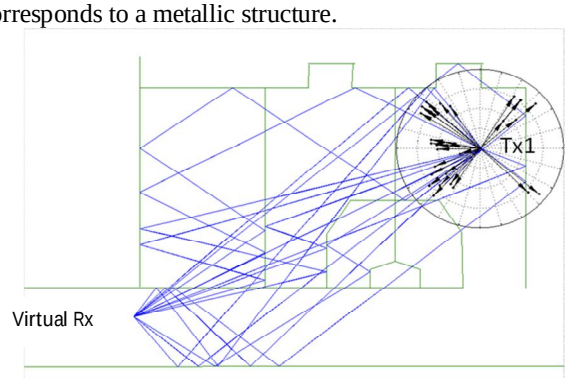


Figure 9. DoD results for the location Tx1.

Fig. 10 presents the comparison between DoD obtained from the measurements and ray-tracing tool for Tx3. As in the previous scenario, a significant agreement is observed between the main directions flows from the measurements and the ray-tracing simulation results. The obtained results also validate our assumption that the propagation in this environment mainly occurs through the corridors (but some waves can propagate through cabin walls). In this configuration, reflections are observed on the corridor walls surrounding the Tx3 location before arriving to the receiver.

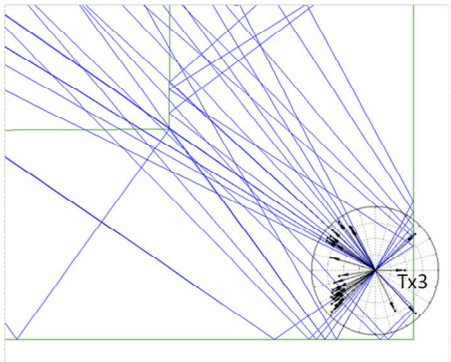


Figure 10. DoD results for the location Tx3.

D. Engineering rules for intermediate nodes placement

As mentioned in Section I, the main objective of these measurements is the optimization of wireless devices placement when deploying wireless networks on board a ship. This study is a part of SAPHIR project, which aims at applying WSN technologies to shipboard alarm and monitoring systems. Sensor nodes must be placed in all ship rooms and compartments to measure physical parameters such as tank level, water level, temperature, humidity, etc. and then send collected data to central control units located in the wheelhouse or the control room. As direct communications between sensor nodes and control units are impossible (due to large distances and metallic environments), intermediate nodes must be located to ensure the whole network connectivity. Results of these measurements are useful for placing intermediate nodes. Therefore, the following engineering rules are recommended. In the engine rooms, EM waves propagate mainly through watertight doors even when they are closed. It will be then recommended to place intermediate nodes in front of watertight doors to ensure connectivity between adjacent rooms. In the parking, the walls constitute a wave guide between communicating nodes. Intermediate nodes may be placed on the walls of the parking, including those of the middle bulkhead, to ensure the network connectivity of different nodes in the parking. In the passenger deck, EM waves propagate mainly through corridors. Intermediate nodes will may be located in the ceiling of corridors (to minimize the fluctuation due to passengers' movement).

IV. CONCLUSION

In this paper, a measurement campaign was conducted on board a ferry using a MIMO channel sounder. The objective of this measurement campaign was to determine the directions of propagation of EM waves in typical shipboard environments. In spite of the totally metallic structure of bulkheads and WT doors in the lowest decks of the ferry, obtained results show that closing WT doors does not block totally the radio waves propagation, which is made through the openings on the edges of the metallic WT doors. Moreover, the results show that parallel metallic bulkheads, such as in the parking, act as a waveguide between Tx and Rx. In spite of NLoS configuration, EM waves are guided through reflections on the metallic bulkheads and diffraction on the bulkheads edges, to the Rx. Measurement and simulation results on the passenger deck show that EM waves propagate mainly through corridors on the passenger deck. When the Tx is located in a passenger cabin, EM waves penetrate through the thin walls of cabins before arriving to the corridor, by which they are guided to the Rx. All these results are used to determine engineering rules for sensor nodes placement.

ACKNOWLEDGMENT

This work is a part of the SAPHIR project supported by the ‘Pôle Mer Bretagne’ and ‘Region Bretagne’. The authors thank Marinelec Technologies and Brittany Ferries for the opportunity to conduct these measurement campaigns.

REFERENCES

[1] J. P. Lynch and K. J. Loh, “A summary review of wireless sensors and sensor networks for structural health monitoring,” The Shock and Vibration Digest, vol. 38, no. 2, 2006, pp.91-128.

[2] D. R. J. Estes, T. B. Welch, A. A. Sarkady, and H. Withesel, “Shipboard radio frequency propagation measurements for wireless networks,” Proceedings of the IEEE Military Communications Conference (MILCOM), IEEE Communications for Network-Centric Operations: Creating the Information Force, Virginia, USA, October 2001, pp. 247-251.

[3] B-G. Paik, S-R. Cho, B-J. Park, D. Lee, B-D. Bae, and J-H. Yun, “Characteristics of wireless sensor network for full-scale ship application,” Journal of Marine Science and Technology, vol. 14, no. 1, January 2009, pp. 115–126.

[4] T. Pilsak, J. L. Ter Haseborg, and H. Hanneken, “WLAN propagation on the bridge of vessels under consideration of materials properties,” Proceedings of the IEEE International Symposium on Electromagnetic Compatibility (EMC Europe), Athens, Greece, June 2009, pp. 1-4.

[5] A. Mariscotti, M. Sassi, A. Qualizza, and M. Lenardon, “On the propagation of wireless signals on board ships,” Proceedings of Instrumentation and Measurement Technology Conference (I2MTC), Austin, Texas, USA, May 2010, pp. 1418-1423.

[6] H. Kdouh, C. Brousseau, G. Zaharia, G. Grunfelder, and G. El Zein, “Measurements and path loss models for shipboard environments at 2.4 GHz,” European Microwave Conference (EuMC), Manchester, United Kingdom, October 2011, pp. 408-411 .

[7] H. Farhat, R. Cosquer, G. Grunfelder, L. Le Coq, and G. El Zein, “A Dual Band MIMO Channel Sounder at 2.2 and 3.5 GHz”, Proceedings of The IEEE International Instrumentation and Measurement Technology Conference (I2MTC 2008), pp. 1980-1985, Victoria, Vancouver Island, Canada, May 2008, pp. 1980-1985.

[8] Y. Lostanlen, H. Farhat, T. Tenoux, A. Carcelen, G. Grunfelder, and G. El Zein, “Wideband Outdoor-to-Indoor MIMO channel measurements at 3.5 GHz”, Proc. of the European Conference on Antennas and Propagation, EUCAP '09, Berlin, Germany, March 2009, pp. 3606-3610.

[9] M. Steinbauer, A.F. Molish, and E. Bonek, “The double directional radio propagation channel,” IEEE Antennas and Propagation Magazine, vol. 43, August 2001, pp. 51-63.

[10] M. S. Bartlett, “An Introduction to Stochastic Process,” New York: Cambridge Univ. Press, 1956.

[11] B. H. Fleury, M. Tschudin, R. Heddergott, D. Dahlhaus, and K. Ingman Pederson, “Channel parameter estimation in mobile radio environments using the SAGE algorithm,” IEEE Journal on Selected Areas in Communications, vol. 17, no. 3, March 1999, pp. 434–450.

[12] Y. Lostanlen, G. Gougeon, S. Bories, and A. Sibille, “A deterministic indoor UWB space-variant multipath radio channel modelling compared to measurements on basic configurations,” Proceedings of the 1st European Conference on Antennas and Propagation, EUCAP ‘06, November 2006, Nice, France, pp. 1-8.

[13] A. Corucci, P. Nepa, F. Furfari, P. Barsocchi, and A. Buffi, “Accuracy limits of in-room localisation using RSSI,” Antennas and Propagation Society International Symposium, APSURSI '09, Pisa, Italy, June 2009, pp. 1-4.

[14] H. Kdouh et al., “Double directional characterisation of radio wave propagation through metallic watertight doors on board ships,” Electronics Letters, vol. 48, no. 6, March 2012, pp. 307-309.

 ISBN: 978-1-61208-279-0

Plasmonics in Optical Communications: Optimization of Coupling Efficiency

Djafar K. Mynbaev
Electrical and Telecommunications Engineering
Technology Department
New York City College of Technology
of the City University of New York
New York, USA
dmynbaev@citytech.cuny.edu

Vitaly Sukharenko
Electrical Engineering Department
City College of New York
of the City University of New York
New York, USA
vsukhar00@ccny.cuny.edu

***Abstract*—Modern optical communications technology requires miniaturization of the end equipment to micro- and even nanoscale. At this scale, interaction of light with this equipment becomes difficult due to diffraction limit. This problem can be solved by the use of plasmonics, which confines light into subwavelength dimension and, additionally, create a strong field enhancement. One of the major plasmonic problems for our application is the efficiency of coupling of light into a metal-dielectric interface, which is mainly determined by the properties of a metal. In this paper, we presenting the results of our search for an optimal (ideal) metal whose properties provide the best coupling efficiency at 1550nm — the main transmission wavelength in optical communications.**

Keywords- Optical communications; plasmonics; coupling efficiency

I. INTRODUCTION

The volume of global telecommunications traffic keeps increasing at the exponential rate and to carry this traffic, the optical networks, which deliver the vast majority of it, exponentially increase its transmission capacity. One of the major technological means to achieve this increase is wavelength-division multiplexing (WDM), in which many transmitters and receivers are used to send and receive many signals over a single optical fiber. Modern WDM technology, however, requires such density of packaging of all components that these components must be in micro- and even nanoscale. Working with light at this scale becomes problematic due to the diffraction limit, the problem that can be overcome by the use of plasmonics.

Plasmonics [1],[2] is about coupling of photons to free electron oscillations at the interface between the thin film of a conductor and a dielectric. This coupling creates two-dimensional electromagnetic waves called surface plasmon polaritons (SPPs) that propagate along this interface. The result is confinement of light into subwavelength dimension, which enables breaking the diffraction limit and creating a strong field enhancement.

Plasmonics has the potential to combine the best properties of both electronic and photonic worlds; in addition, plasmonics allows for reducing light manipulation from three to two dimensions. All these features might lead to creation of integrated photonic circuits, in which optical communications is in dire need.

From the optical-communications standpoint, the most important feature of plasmonics is that the SPPs can be seen as a new optical carrier of information that allows signal manipulation at the scale below diffraction limit. This is why application of plasmonics in optical communications has attracted a significant interest of a research community [3]–[5]. One of the main problems of application of plasmonics in optical communications is the efficiency of coupling of light into a metal because of tremendous loss that light experiences in a metal film. Clearly, the coupling efficiency shapes the total efficiency of any plasmonic device. The coupling efficiency is determined by the properties of the metal; for visible light, which is mostly used in current research, the best coupling has been achieved with silver and gold due to their beneficial permittivity in this range of wavelengths. Optical communications, however, operates in infrared segment of the spectrum, and the range of wavelengths around 1550nm is most widely used. There are no experimental or theoretical results indicating which metal could provide the optimum coupling in this range. In this work, we present the results of our search for an optimal material providing the best possible coupling efficiency at 1550nm.

The rest of this paper is organized as follows: In Section II, we evaluate the coupling efficiency of the metals mostly used in plasmonics and find their optimal thickness; in Section III, we apply Occam inversion method to find the most efficient optimal material; in Section IV, we present the summary and directions of the future work.

II. COUPLING EFFICIENCY AND OPTIMAL THICKNESS OF AL, CU, PB, AU, AND AG

In this section, we investigate plasmonic properties of Al, Cu, Pb, Au and Ag, the metals mostly used in experimental research. Some of their properties have been studied for visible light; we, however, need to know these properties at the main transmission wavelength of optical communications, $\lambda = 1550$nm.

Figure 1 illustrates how light is coupled at the metal-dielectric interface; to optimize this process, we need to achieve the maximum penetration of light from the outside

 ISBN: 978-1-61208-279-0

(not shown) material into a metal and obtain the maximum energy of the filed penetrated into a dielectric. Penetration of light into the metal is evaluated by reflection and transmission coefficients; the latter determines the strength of a resonance field, the field penetrated into a dielectric. This strength is crucial because it determines the length the excited light can propagate within the dielectric, which in turn determines the efficiency of the whole plasmonic process in our application.

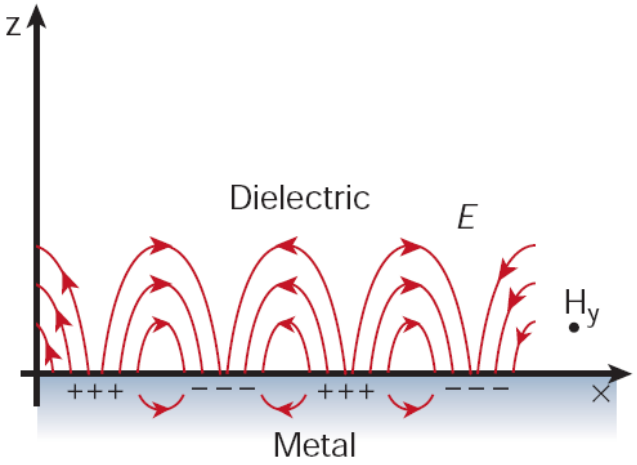


Figure 1. Coupling of light in plasmonics: Electron charges and excited electric waves. (Reprinted with permission [6])

In this section, we consider the setup composed from three media: glass-metal-air. The refractive indexes of glass and air are given by n_g = 1.5 and n_a = 1, respectively. For this setup we determine the coupling efficiency by finding the reflection and transmission coefficients for arbitrary thickness of each metal. Analysis of these results leads us to the necessity of search of optimal thickness of a metal, which we do as follows.

The Drude model [7] defines dielectric function of free electron plasma and the electric permittivity, ε, of a metal to the optical response as

$$\varepsilon = \varepsilon_\infty - \frac{\omega_p^2}{\omega^2 + i\omega\omega_\tau} \quad (1)$$

where ω_p is the plasma, ω is the incident, and ω_τ is the damping frequencies and ε_∞ is the high-frequency dielectric constant. Using the values of plasma frequency and damping factor from the study conducted by Ordel [7], we calculate the electric permittivity of each metal at λ = 1550nm; the results are shown in Table I.

Using Maxwell's equations and appropriate boundary conditions for our setup, one can derive the formulas (Equations 2, 3 and 4 [2]) that allow for evaluation of reflection and transmission of each metal as a function of incident angle.

TABLE I. ELECTRIC PERMITIVITY OF ALUMINUM, COPPER, LEAD, GOLD AND SILVER

Material	Permittivity at 1550nm
Al	-335.83+33.779i
Cu	-96.6111 + 4.2061i
Pb	-91.3055 + 2.0746i
Au	-125.14+4.2232i
Ag	-125.22+2.8367i

$$R_{gm} = \frac{\varepsilon_m k_g(\theta_g) - \varepsilon_g k_m(\theta_g)}{\varepsilon_m k_g(\theta_g) + \varepsilon_g k_m(\theta_g)} \quad (2)$$

$$R_{ma} = \frac{\varepsilon_a k_m(\theta_g) - \varepsilon_m k_a(\theta_g)}{\varepsilon_a k_m(\theta_g) + \varepsilon_m k_a(\theta_g)} \quad (3)$$

$$R_{gma} = \frac{R_{gm} + R_{ma}\exp(i2k_m d_m)}{1 + R_{gm}R_{ma}\exp(i\,2k_m d_m)}, \quad (4)$$

where $\varepsilon_g, \varepsilon_m, \varepsilon_a$ are the electric permittivity's of glass, metal and air, respectively; k_g, k_m, k_a are the wave numbers of glass, metal, and air; d_m is the metal thickness. We have taken electric permittivity of each metal from Table 1 and perform simulations.

Plots in Figure 2 present the results of our calculations of reflection and transmission coefficients of all mentioned metals at arbitrary 55nm thickness for each of them. Our results show that the thickness of a metal is the second major factor determining the coupling efficiency. We investigated this factor for each material; the samples of such a search for gold are shown in Figures 3 and 4. This example shows that the minimum reflection and the maximum transmission for gold are achieved at 35nm.

Based on the results of our search, we calculated the reflection and transmission coefficients for each metal at their appropriate optimal thickness. Figure 5 show the coupling efficiency of 70% for aluminum at 20nm thickness, 100% for copper at 40nm thickness, 95% for lead at 48nm thickness, 95% for gold at 34nm thickness, and 80% for silver at 32nm thickness.

Comparison of the intensity of the field penetrated into a dielectric (see Figure 5, transmission) shows that the strength of a penetrated field varies significantly for each metal; for example, silver exhibits the strongest resonance field, 375 a.u., at slightly less than 80% coupling (Figure 5, reflection), whereas lead exhibits 100% of coupling, but its resonance field is only 25 a.u. strong. Therefore, when searching a metal, which provides the best coupling, we need to consider its reflection and transmission coefficients at its optimal thickness. Only combination of all three parameters provides the optimal coupling efficiency evaluated by reflection, transmission, and the strength of resonance (penetrated) field.

 ISBN: 978-1-61208-279-0

The summary of these results are presented in Table II.

TABLE II. STRENGTH OF A PENETRATED FIELD FOR ALUMINUM, COPPER, LEAD, GOLD AND SILVER

Material	Thickness	Field Strength	Coupling
Al	20nm	20a.u.	70%
Cu	40nm	55a.u.	100%
Pb	48nm	60a.u.	95%
Au	34nm	120a.u	95%
Ag	32nm	375a.u.	80%

III. OCCAM INVERSION AND IDEAL (DESIRED) MATERIAL

In this section, we describe our search of electric permittivity of an ideal material that has high coupling ratio and generates strong resonances on its surface due to the incident radiation at 1550nm. We consider the setup composed from three media, as in the previous section, but instead of the metal we place an ideal material. Using Occam inversion technique [8] and Maxwell's equation with appropriate boundary conditions, we can calculate electric permittivity of a material that generates high resonances on its surface and satisfies above criteria: 100% coupling and long range propagation of excited field.

Occam inversion algorithm, (5), searches for optimal electric permittivity by employing angular scattering, R_{gma}, (4). The algorithm minimizes the difference between the modeled, R_{gma}, and a desired scattering pattern. With each iteration, the obtained solution, ε^{k+1}, is closer to the optimal permittivity; the process stops when the difference between the obtained and previous solutions becomes less than 10^{-9}. Desired scattering pattern exhibits 100% coupling at a given metal thickness. For these calculations we use least-squares method and assume 65nm metal thickness.

$$\varepsilon^{k+1} = [J(\varepsilon^k)^T J(\varepsilon^k) + \alpha^2 L^T L]^{-1} J(\varepsilon^k)^T R. \quad (5)$$

Here $J(\varepsilon^k)$ is the Jacobian of $G(\varepsilon^k)$, L - laplacian operator, α is the regularization parameter, and R is the desired reflection pattern. Using initial estimation ε^0, which can be arbitrarily chosen, we found the following ideal permittivity

$$\varepsilon_{ideal} = -121.677 + 0.5243019\mathrm{i}\ . \quad (6)$$

It is worth mentioning that choosing the initial estimation, ε^0, closer to the perfect value will essentially reduce the number of iterations and minimize the simulation process.

Reflection and transmission patterns of an ideal material at thickness 65nm and that of Ag and Pb at their optimal thicknesses plotted in Figure 6. Ag and Pb are chosen as control metals to contrast reflected and transmitted patterns of the optimal material; Ag has highest transmitted patterns at optimal thickness and Pb requires larger thickness for optimal coupling.

The optimal material shows the maximum coupling (0% reflection), but very weak penetrated field, about ten times smaller than that of silver. Thus, there is a need to optimize the thickness of this metal. The results of this search are presented in Figures 7 and 8, where we plot all reflection and transmission coefficients of above mentioned metals at 55nm thickness (Figure 7) and at 35nm thickness (Figure 8). One can clearly see that at 55nm an optimal material has about 85% of coupling efficiency but the strength of penetrated field is only 160 a.u., whereas at 35nm this material exhibits almost 7000 a.u. of the resonance field at the expense of coupling, which is reduced to 30%. Thus, these two processes contradict each other. Recalling that for ideal material we need 100% of coupling (0% of reflection) and the highest energy of penetrated field (e.g., 10,000 a.u.) to provide the longest propagation of the excited field in a dielectric, we come to the conclusion that required optimization is not an easy task.

IV. SUMMARY AND CONCLUSION

We can summarize the results presented in this paper as follows:

1. We have found out the reflection coefficients, the strengths of resonance fields and optimal thickness of a set of natural metals widely used in plasmonics. To our knowledge, we did this the first time at λ = 1550nm, the main transmission wavelength in optical communications. The importance of these results cannot be overestimated they have determined the fundamental limitations imposed on the application of these metals in plasmonic devices in optical communications.
2. We have searched for the desired (optimal) material, which could provide the optimal coupling efficiency. We have determined its desired characteristics+, such as electric permittivity, reflection coefficient and the strength of a resonance field; we have proven that this material could achieve the increase of the strength of its field by an order of magnitude at 1550nm , as compared with Ag and Pb.

Our results lead us to the following directions of the future work:

1. Considering the characteristics of the desired (optimal) material, we note that the real part of its electric permittivity is close to that of silver and gold, but its imaginary part is tenfold less than the proper parts of these natural metals. It means that such a desired (optimal) material does not exist in nature and we need to artificially create a material with the required electric permittivity. This is one direction of the future work.
2. Our simulations have shown that the requirements of simultaneously maximizing both the coupling and the strength of penetrated field contradict each other. Therefore, we need to either optimize the properties of the desired material or find the different approach

 ISBN: 978-1-61208-279-0

to satisfy these mutually contradictory requirements. This work would be the other direction of our future research.

3. We will need to find out how our desired (optimal) material would work not only at a single 1550nm wavelength, but also in both C- and L-bands used in modern DWDM optical transmission networks.

ACKNOWLEDGMENT

This work is partly supported by PSC-CUNY grant #65115-00 43.

REFERENCES

[1] S. A. Maier, "Plasmonics: Fundamentals and Applications," Springer Science+Business Media LLC, 2007.

[2] D. Sarid and W. Challener, "Modern Introduction to Surface Plasmons," Cambridge University Press, 2010.

[3] H. A. Atwater et al, "Novel Plasmonic Devices for Nanophotonic Networks," See www.plasmonmuri.caltech.edu/news/PosterAFOSR.pdf, 2011 (accessed February 2, 2013).

[4] D. K. Mynbaev and V. Sukharenko, "WDM Demultiplexing by Using Surface Plasmon Polaritons," International Journal of High Speed Electronics and Systems, Vol. 20, No. 1, March 2011, Pp. 51-61.

[5] D. K. Mynbaev and V. Sukharenko, "Plasmonic-based devices for optical communications," International Journal of High Speed Electronics and Systems, Vol. 21, No. 1, March 2012.

[6] I. Zozoulenko, "Surface plasmons and their applications in electro-optical devices," Solid State Electronics, Department of Science and Technology, Linköping University, Sweden, 2006. See http://www.itn.liu.se/meso-phot (accessed January 12, 2013).

[7] A. Ordal et al, "Optical properties of the metals Al, Co, Cu, Au, Fe, Pb, Ni, Pd, Pt, Ag, Ti and W in the infrared and far infrared," Applied Optics Vol. 22, No. 7 1, April 1983.

[8] R. C. Aster, B. Borchers, and C. H. Thurber, "Parameter Estimation and Inverse Problems," Elsevier Academic Press, 2005.

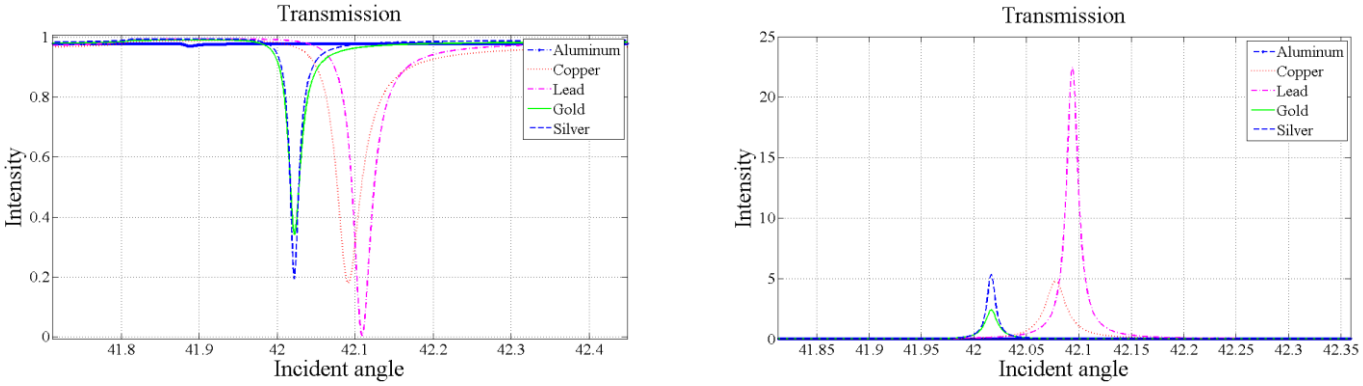


Figure 2. Reflection and transmission patterns of aluminum, copper, lead, gold and silver at equal thicknesses of 55nm.

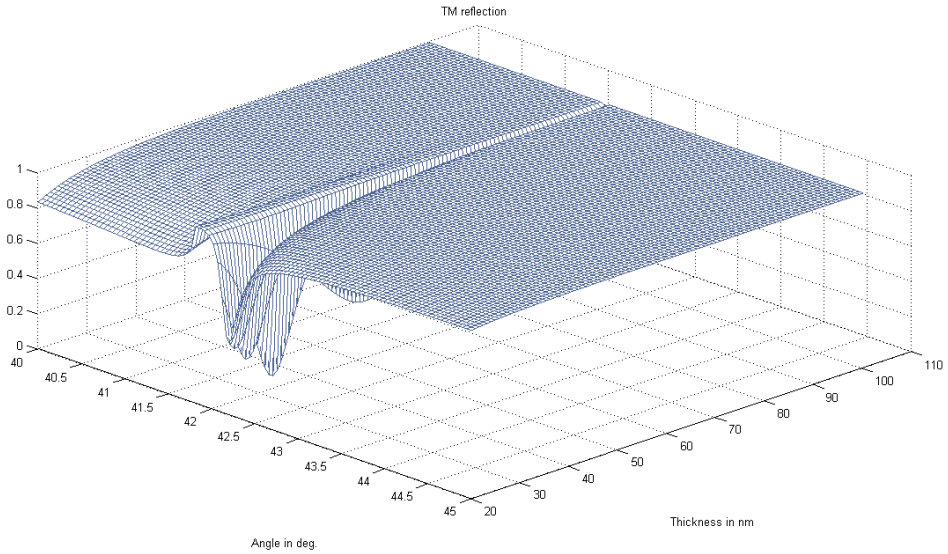


Figure 3. Optimal thickness of gold providing the minimum of reflection.

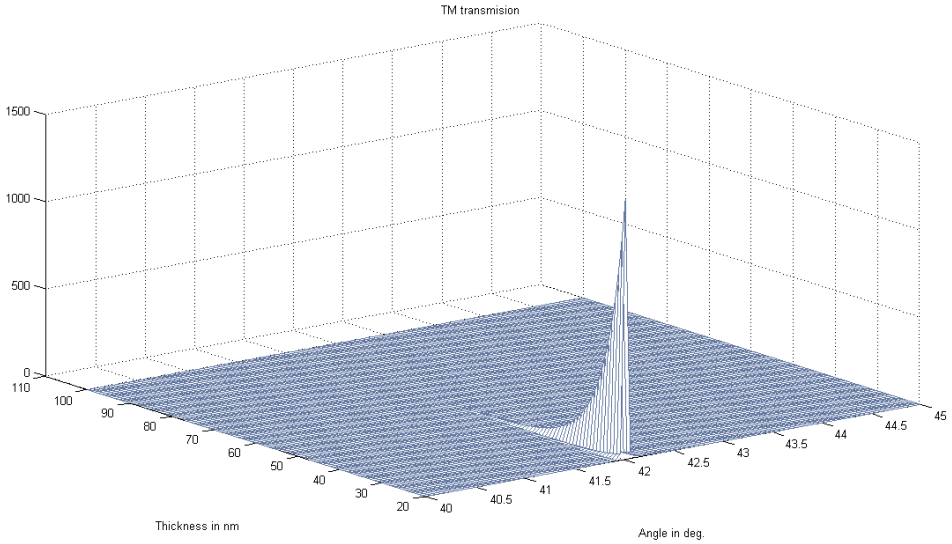


Figure 4. Optimal thickness of gold providing the maximum of transmission.

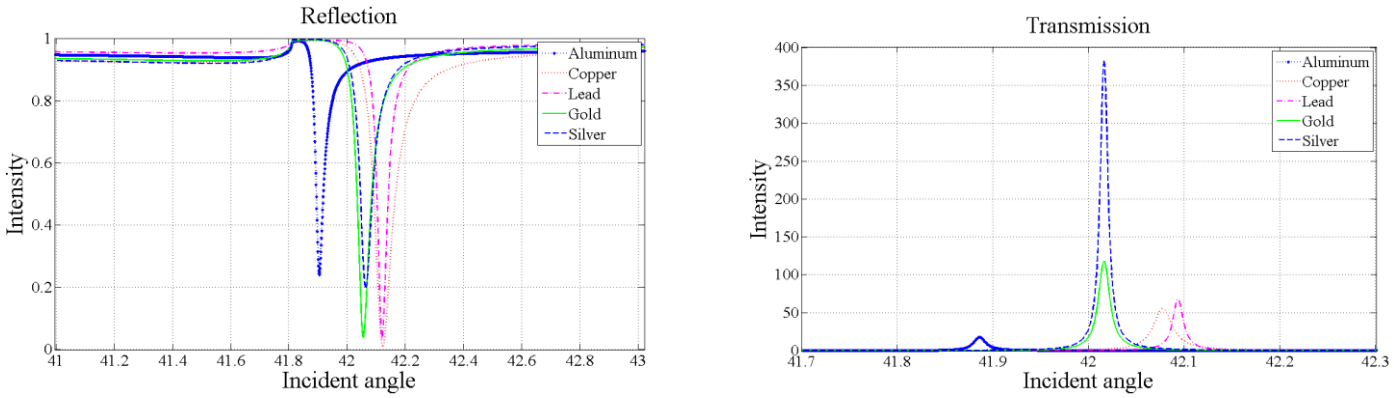


Figure 5. Reflection and transmission patterns of aluminum at thickness of 20nm, copper at thickness of 40nm, lead at thickness of 48nm, gold at thickness of 34nm, and silver at thickness of 32nm.

 ISBN: 978-1-61208-279-0

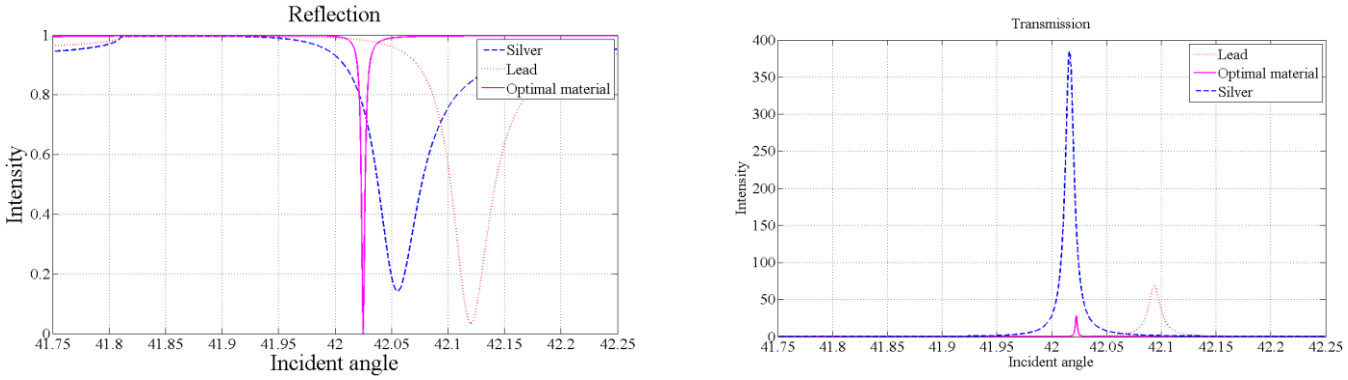


Figure 6. Reflection and transmission patterns of lead at thickness of 48nm, silver at thickness of 34nm and optimal material at thickness of 65nm.

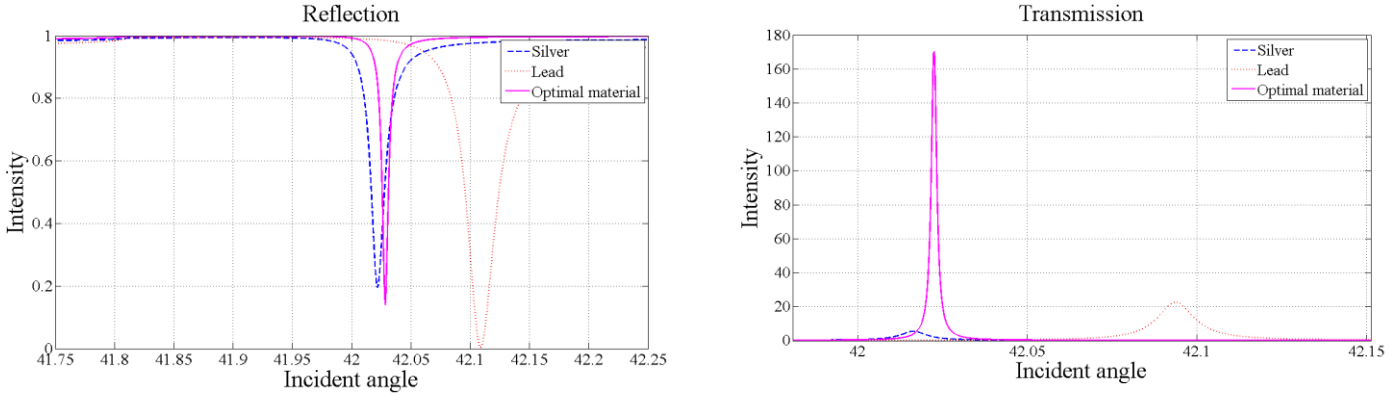


Figure 7. Reflection and transmission patterns of silver, lead, and optimal material at thicknesses of 55nm.

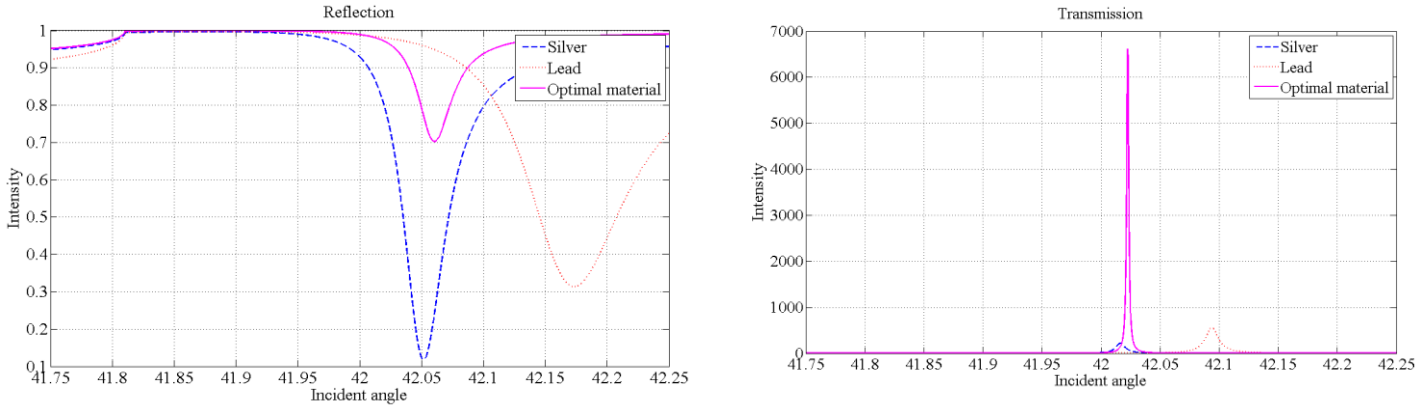


Figure 8. Reflection and transmission patterns of silver, lead, and optimal material at thicknesses of 35nm.

 ISBN: 978-1-61208-279-0