ICWMC 2012

The Eighth International Conference on Wireless and Mobile Communications

ISBN: 978-1-61208-203-5

June 24-29, 2012

Venice, Italy

ICWMC 2012 Editors

Dragana Krstic, University of Nis, Serbia

Eugen Borcoci, University “Politehnica” Bucharest, Romania

ICWMC 2012

Foreword

The Eighth International Conference on Wireless and Mobile Communications [ICWMC 2012], held between June 24-29, 2012 - Venice, Italy, followed on the previous events on advanced wireless technologies, wireless networking, and wireless applications.

ICWMC 2012 addressed wireless related topics concerning integration of latest technological advances to realize mobile and ubiquitous service environments for advanced applications and services in wireless networks. Mobility and wireless, special services and lessons learnt from particular deployment complemented the traditional wireless topics.

We take here the opportunity to warmly thank all the members of the ICWMC 2012 Technical Program Committee, as well as the numerous reviewers. The creation of such a 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 ICWMC 2012. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

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

We hope that ICWMC 2012 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the area of wireless and mobile communications.

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

ICWMC 2012 Chairs:

Dragana Krstic, University of Nis, Serbia
Magnus Jonsson, Halmstad University, Sweden
Yan Zhang, IMEC- NL, The Netherlands
Christopher Nguyen, Intel Corp., USA
Georg Neugebauer, RWTH Aachen University, Germany

ICWMC 2012

Committee

ICWMC Advisory Committee

Dragana Krstic, University of Nis, Serbia
Magnus Jonsson, Halmstad University, Sweden

ICWMC Industry/Research Chairs

Yan Zhang, IMEC- NL, The Netherlands
Christopher Nguyen, Intel Corp., USA

ICWMC Special Area Chairs

Security
Georg Neugebauer, RWTH Aachen University, Germany

ICWMC 2012 Technical Program Committee

Jemal Abawajy, Deakin University - Victoria, Australia
Mohammed Abdel-Hafez, UAE University-Al-Ain, United Arab Emirates
Seyed Reza Abdollahi, Brunel University - London, UK
Javier M. Aguiar Pérez, Universidad de Valladolid, Spain
Mari Carmen Aguayo-Torres, Universidad de Malaga, Spain
Chang-Jun Ahn, Chiba University, Japan
Ahmed Akl, LAAS/CNRS - Toulouse, France
Hamed Al-Raweshidy, Brunel University - Uxbridge, UK
Erick Amador, Intel Mobile Communications - Sophia Antipolis, France
Radu Arsinte, Technical University of Cluj-Napoca, Romania
Ruzena Bajcsy, University of California - Berkeley, USA
Mohammad M. Banat, Jordan University of Science and Technology, Jordan
Alessandro Bazzi, IEIIT-CNR, University of Bologna, Italy
Norman C. Beaulieu, University of Alberta, Canada
Mouncef Benmimoune, Universite du Quebec a Trois-Rivieres, Canada
Ezio Biglieri, Universitat Pompeu Fabra - Barcelona, Spain
Alireza Borhani, University of Agder - Grimstad, Norway
David Boyle, Tyndall National Institute, University College Cork, Ireland
Maurizio Bozzi, University of Pavia, Italy
Maria Calderon, University Carlos III of Madrid, Spain
Juan-Carlos Cano, Universidad Politécnica de Valencia, Spain
Pedro Castillejo Parrilla, Universidad Politécnica de Madrid, Spain
Sammy Chan, City University of Hong Kong, Hong Kong
Ajit Chaturvedi, Indian Institute of Technology Kanpur, India

Timo O. Korhonen, Aalto University, Finland
Dragana Krstic, University of Nis, Serbia
Alexey Lagunov, M.V. Lomonosov - The Northern (Arctic) Federal University, Russia
Zhihua Lai, Ranplan Wireless Network Design Ltd., UK
Jingli Li, TopWorx - Emerson, USA
Qilian Liang, Wuhan University, China
Fidel Liberal Malaina, University of the Basque Country, Spain
Justin Lipman, Intel R&D China, China
Donggang Liu, Wuhan University, China
Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Andreas Löffler, Friedrich-Alexander University of Erlangen-Nüremberg, Germany
Valeria Loscrí, University of Calabria, Italy
Jonathan Loo, Middlesex University - London, UK
Lin Luo, University of South Australia - Mawson Lakes, Australia
Christian Maciocco, Intel Corporation -Santa Clara, USA
D. Manivannan (Mani), University of Kentucky - Lexington, USA
Muneer Masadeh Bani Yassein, Jordan University of Science and Technology - Irbid, Jordan
Barbara M. Masini, CNR - IEIIT, University of Bologna, Italy
Daniel Massicotte, Université du Québec à Trois-Rivières, Canada
Catherine Meadows, Naval Research Laboratory - Washington DC, USA
Hamid Menouar, QUWIC - Qatar University Wireless Innovation Center, Qatar
Makoto Miyake, M-TEC Company Limited / Mitsubishi Electric Corporation, Kamakura-City, Japan
Klaus Moessner, University of Surrey, UK
Mohamed Moustafa, Akhbar El Yom Academy, Egypt
Lorenzo Mucchi, University of Florence, Italy
Renato Negra, RWTH Aachen University, Germany
Marek Neruda, Czech Technical University in Prague, Czech Republic
Georg Neugebauer, RWTH Aachen University, Germany
Christopher Nguyen, Intel Corp., USA
Homayoun Nikookar, Delft University of Technology, The Netherlands
Loutfi Nuaymi, Telecom Bretagne - Rennes, France
Shigeaki (Aki) Ogose, Kagawa University, Japan
George S. Oreku, TIRDO/ North west University, South Africa
Tudor Palade, Technical University of Cluj-Napoca, Romania
Carlos Enrique Palau Salvador, Polytechnic University of Valencia, Spain
Salvatore Flavio Pileggi, Universidad Politécnica de Valencia, Spain
Thomas Plos, TU-Graz, Austria
Anastasios Politis, Technological Educational Institute of Serres, Greece
Carlos Pomalaza-Raez, Purdue University, USA / University of Oulu, Finland
Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland
Anand R. Prasad, NEC Corporation, Japan
Hani Ragab Hassen, University of Greenwich, UK
Yusnita Rahayu, Universiti Malaysia Pahang (UMP), Malaysia
Teng Rui , NICT, Japan
José Antonio Sánchez Fernández, Universidad Politécnica de Madrid, Spain
David Sánchez Rodríguez, University of Las Palmas de Gran Canaria (ULPGC), Spain
José Santa Lozano, Universidad de Murcia, Spain
Reijo Savola, VTT, Finland

Hans-Otto Scheck, Nokia Siemens Networks, Germany
Riccardo Scopigno, Istituto Superiore Mario Boella - Torino, Italy
Zary Segall, Royal Institute of Technology (KTH), Sweden
Jean-Pierre Seifert, Technische Universität Berlin & Deutsche Telekom Laboratories - Berlin, Germany
Sandra Sendra Compte, Polytechnic University of Valencia, Spain
Ali Shahrabi, Glasgow Caledonian University, UK
Adão Silva, University of Aveiro / Institute of Telecommunications, Portugal
Wojciech Siwicki, Gdansk University of Technology, Poland
Mariusz Skrocki, Orange Labs - Warszawa, Poland
Vahid Solouk, Urmia University of Technology, Iran
Himanshu B Soni, G.H. Patel College of Engineering & Technology, India
Kuo-Feng Ssu, National Cheng Kung University, Taiwan
Razvan Stanica, National Polytechnic Institute of Toulouse, France
Petr Svenda, Masaryk University - Brno, Czech Republic
Yasihisa Takizawa, Kansai University, Japan
Fatma Tansu Hocanin, Eastern Mediterranean University, Turkey
Thomas Ußmüller, University of Erlangen-Nuremberg, Germany
Václav Valenta, Ulm University, Germany
K. Vasudevan, Indian Institute of Technology - Kanpur, India
Natalija Vlajic, York University - Toronto, Canada
Baptiste Vrigneau, IUT - R&T Châtellerault, France
Robert Weigel, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Yean-Fu Wen, National Chiayi University, Taiwan
Martin Werner, Ludwig-Maximilians-University Munich, Germany
Ouri Wolfson, University of Illinois at Chicago, USA
Qishi Wu, University of Memphis, USA
Erkan Yüksel, Istanbul University, Turkey
Sherali Zeadally, University of the District of Columbia, USA
Yuanyuan Zeng, Wuhan University, China
Hans-Jürgen Zepernick, Blekinge Institute of Technology, Sweden
Yan Zhang, IMEC- NL, The Netherlands

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

Performance Evaluation of Multicell Coordinated Beamforming Approaches for OFDM Systems

José Assunção, Reza Holakouei, Adão Silva, and Atílio Gameiro
DETI, Instituto de Telecomunicações, University of Aveiro, Portugal
E-mails: jassuncao@av.it.pt ,rholakouei@ua.pt , asilva@av.it.pt, and amg@ua.pt

Abstract - In this paper we propose and evaluate multicell coordinated beamforming schemes for the downlink of MISO-OFDM systems. The precoders are designed in two phases: first the precoder vectors are computed in a distributed manner at each BS considering two criteria, namely distributed zero-forcing and virtual signal-to-interference noise ratio. Then the system is optimized through distributed power allocation under per-BS power constraint. The proposed power allocation scheme is designed based on minimization of the average bit error rate over all the available subcarriers. Both the precoder vectors and the power allocation are computed by assuming that the BSs have only knowledge of local channel state information and do not share the data symbols. The performance of the proposed schemes are evaluated, considering typical pedestrian scenarios based on LTE specifications. The results have shown that the proposed distributed power allocation scheme outperform the equal power allocation approach.

Keywords-component; distributed precoding, distributed power allocation, multicell systems, OFDM and LTE.

I. INTRODUCTION

Multicell cooperation is one of the fastest growing areas of research, and it is a promising solution for cellular wireless systems to mitigate intercell interference, improving system fairness and increasing capacity in the years to come. This technology is already under study in LTE-Advanced under the coordinated multipoint (CoMP) concept.

There are several CoMP approaches depending on the amount of information shared by the transmitters through the backhaul network and where the processing takes place, i.e., centralized if the processing takes place at the central unit (CU) or distributed if it takes at the different transmitters. Coordinated centralized beamforming approaches, where transmitters exchange both data and channel state information (CSI) for joint signal processing at the CU, promise larger spectral efficiency gains than distributed interference coordination techniques, but typically at the price of larger backhaul requirements and more severe synchronization requirements. Two centralized multicell precoding schemes based on the waterfilling technique have been proposed in [1]. It was shown that these techniques achieve a performance, in terms of weighted sum rate, very close to the optimal. In [2] a clustered BS coordination is enabled through a multicell block diagonalization (BD) strategy to mitigate the effects of interference in multicell MIMO systems. A new BD cooperative multicell scheme has been proposed in [3], to maximize the weighted sum-rate achievable for all the user terminals (UTs).

Distributed precoding approaches, where the precoder vectors are computed at each BS in a distributed fashion, have been proposed in [4]. It is assumed that each base station has only the knowledge of local CSI and based on that a parameterization of the beamforming vectors used to achieve the outer boundary of the achievable rate region was derived. In [5], distributed precoding schemes based on zero-forcing criterion with several centralized power allocation based on minimization of the average BER and sum of inverse of signal-to-noise ratio (SNIR) have been derived.

In the previous approaches, it was assumed that the transmitters (or BSs) share the entire data of all UTs. However, there are distributed beamforming approaches where the transmitters do not share the data, which fall into the interference channel (IC) framework. The local CSI, i.e. the CSI between a given BS and all UTs, is used by transmitters to design individual precoders to transmit exclusively to the users within their own cell [6], [7]. This approach, known as inter-cell interference nulling (ICIN), in which each BS transmits in the null-space of the interference it is causing to neighboring cells, has been discussed in the 3GPP long term evolution advanced (LTE-A) literature. The authors of [8] proposed a non-iterative distributed solution to design precoding matrices for multi-cell systems, which maximizes the sum-rates for only a two-cell system at high SNR. In [9], a coordinated beamforming approach based on the virtual SINR framework, for a special case of two transmitters, has been proposed.

The aim of this work is to propose and evaluate coordinated beamforming for the downlink of multicell MISO-OFDM systems. It is assumed that the BSs have only knowledge of local CSI and do not share the data symbols. The precoder is designed in two phases: first the precoder vectors are computed based on distributed zero-forcing (DZF), and distributed virtual signal-to-interference noise ratio (DVSINR). Then the system is further optimized by proposing a novel distributed power allocation algorithm, based on minimization of the average bit error rate (BER) over the available subcarriers. With the proposed strategy both the precoder vectors and the power allocation are computed at each BS in a distributed manner. The considered criterion for power allocation essentially lead to a redistribution of powers among subcarriers, and therefore

 ISBN: 978-1-61208-203-5

provide data symbols fairness, which in practical cellular systems may be for the operators a goal as important as throughput maximization.

The remainder of the paper is organized as follows: section II presents the multicell MISO-OFDM system model. Section III briefly describes the considered distributed precoder vectors. In Section IV the novel distributed power allocation scheme is derived. Section V presents the main simulation results. The conclusions will be drawn in section VI.

II. System Model

Throughout this paper, we will use the following notations. Lowercase letters, boldface lowercase letters and boldface uppercase letters are used for scalars, vectors and matrices, respectively. $(.)^H$ represents the conjugate transpose operators, E[.] represents the expectation operator, $\mathbf{I}_N$ is the identity matrix of size $N \times N$, $\mathcal{CN}(.,.)$ denotes a circular symmetric complex Gaussian vector and χ_n^2 denotes the chi-square random variable with *n* degrees of freedom.

We consider the MISO interference channel where B BSs, each equipped with N_{t_b} antennas, transmit to B single antenna UTs, as shown in Fig. 1. Also, we assume an OFDM based system with N_c available subcarriers. Under the assumption of linear precoding, the signal transmitted by the BS *b* on subcarrier l is given by,

$$\mathbf{x}_{b,l} = \sqrt{p_{b,l}}\,\mathbf{w}_{b,l} s_{b,l} \tag{1}$$

where $p_{b,l}$ represents the transmitted power allocated to sub-carrier l at BS b, $\mathbf{w}_{b,l} \in \mathrm{C}^{N_{t_b} \times 1}$ is the precoder at BS b on sub-carrier l with unit norms, i.e., $\|\mathbf{w}_{b,l}\| = 1,\ b = 1,...,B,\ l = 1,...,N_c$. The data symbol $s_{b,l}$, with $\mathrm{E}\left[\left|s_{b,l}\right|^2\right] = 1$, is intended for UT b. The average power transmitted by the BS b is then given by,

$$\mathrm{E}\left[\left\|\mathbf{x}_b\right\|^2\right] = \sum_{l=1}^{N_c} p_{b,l} \tag{2}$$

where $\mathbf{x}_b$ is the signal transmitted over the N_c subcarriers.

The received signal at the UT b on sub-carrier l, $y_{b,l} \in \mathrm{C}^{1\times 1}$, can be expressed by,

$$y_{b,l} = \sum_{j=1}^{B} \sqrt{p_j}\,\mathbf{h}_{j,b,l}^H \mathbf{w}_{j,l} s_{j,l} + n_{b,l} \tag{3}$$

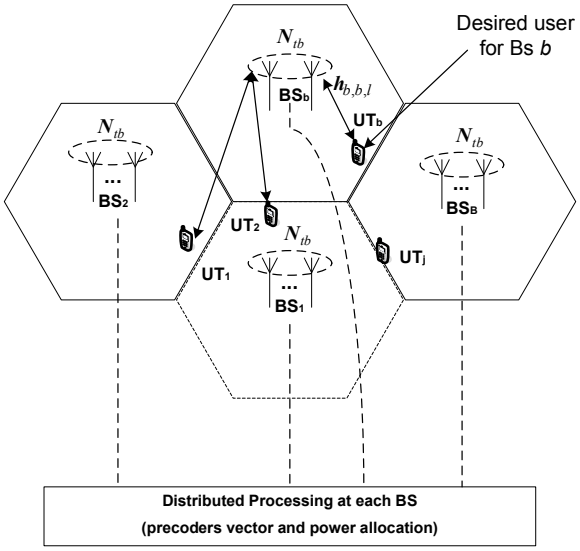


Fig. 1: System Model considered.

where $\mathbf{h}_{j,b,l} \sim \mathcal{CN}\left(0, \rho_{j,b}\mathbf{I}_{N_{t_b}}\right)$ of size $N_{t_b} \times 1$, represents the channel between user b and BS j on subcarrier l and $\rho_{j,b}$ is the long-term channel power gain between BS j, and UT b and $n_{b,l} \sim \mathcal{CN}\left(0, \sigma^2\right)$ is the noise.

From (1) and (3) the received signal at UT b on sub-carrier l can be decomposed in,

$$y_{b,l} = \underbrace{\sqrt{p_{b,l}}\,\mathbf{h}_{b,b,l}^H \mathbf{w}_{b,l} s_{b,l}}_{\textit{Desired Signal}} + \underbrace{\sum_{\substack{j=1 \\ j\neq b}}^{B} \sqrt{p_{j,l}}\,\mathbf{h}_{j,b,l}^H \mathbf{w}_{j,l} s_{j,l}}_{\textit{Multiuser Multicell Interference}} + \underbrace{n_{b,l}}_{\textit{Noise}} \tag{4}$$

and from (4) the instantaneous SINR of user b on sub-carrier l can be written as,

$$\mathrm{SINR}_{b,l} = \frac{\left|\sqrt{p_{b,l}}\,\mathbf{h}_{b,b,l}^H \mathbf{w}_{b,l}^{(type)}\right|^2}{\sum_{\substack{j=1 \\ j\neq b}}^{B} \left|\sqrt{p_{j,l}}\,\mathbf{h}_{j,b,l}^H \mathbf{w}_{j,l}^{(type)}\right|^2 + \sigma^2} \tag{5}$$

where type = {DZF,DVSINR}. Assuming M-ary QAM constellations, the instantaneous probability of error for user b and data symbol transmitted on subcarrier l is given by [10],

$$P_{e,b,l} = \psi Q\left(\sqrt{\beta SINR_{b,l}}\right) \tag{6}$$

where $Q(x) = \left(1/\sqrt{2\pi}\right)\int_x^\infty e^{-\left(t^2/2\right)} dt$, $\beta = 3/\left(M-1\right)$ and $\psi = \left(4/\log_2 M\right)\left(1-1/\sqrt{M}\right)$.

III. DISTRIBUTED PRECODER VECTORS

In this section we describe the distributed precoding vectors, namely DZF and DVSINR. To design the distributed precoder vector we assume that the BSs have only knowledge of local CSI and its own data symbols, i.e., BS b knows the instantaneous channel vectors $\mathbf{h}_{b,j,l}, \forall j,l$, and only the data symbols $s_{b,l}, l = 1,\ldots,N_c$ reducing the feedback load over the backhaul network as compared with the data and/or CSI sharing beamforming approaches.

A. Distributed Zero Forcing (DZF)

Zero forcing is considered a classic beamforming strategy which removes the co-terminal interference. We derive a distributed ZF transmission scheme with the phase of the received signal at each UT aligned. In this case, $\mathbf{w}_{b,l}^{(DZF)}$ in (5) is a unit-norm zero forcing vector orthogonal to $B-1$ channel vectors $\left\{\mathbf{h}_{b,j,l}^H\right\}_{j\neq b}$. By using such precoding vectors, the multicell interference is canceled and the data symbol at each BS on each subcarrier is only transmitted to its intended UT. The SVD of $\left\{\mathbf{h}_{b,j,l}^H\right\}_{j\neq b}$ can be portioned as follows,

$$\left\{\mathbf{h}_{b,j,l}^H\right\}_{j\neq b} = \mathbf{U}_{b,l}\Omega_{b,l}\left[\ddot{\mathbf{W}}_{b,l} \;\; \bar{\mathbf{W}}_{b,l}\right] \tag{7}$$

where $\bar{\mathbf{W}}_{b,l} \in \mathbf{C}^{N_{t_b}\times\left(N_{t_b}-B+1\right)}$ holds the $\left(N_{t_b}-B+1\right)$ singular vectors in the null space of $\left\{\mathbf{h}_{b,j,l}^H\right\}_{j\neq b}$.The columns of $\bar{\mathbf{W}}_{b,l}$ are candidates for b's precoding vector since they will produce zero interference at the other UTs. It can be shown that an optimal linear combination of these vectors can be given by [5],

$$\mathbf{w}_{b,l}^{(DZF)} = \bar{\mathbf{W}}_{b,l}\frac{\left(\mathbf{h}_{b,b,l}^H\bar{\mathbf{W}}_{b,l}\right)^H}{\left\|\mathbf{h}_{b,b,l}^H\bar{\mathbf{W}}_{b,l}\right\|} \tag{8}$$

Also, it can be shown that $\mathbf{h}_{b,b,l}^H\mathbf{w}_{b,l}^{(DZF)} \sim \chi^2_{2\left(N_{t_b}-K+1\right)}$.

B. Distributed Virtual SINR (DVSINR)

Intuitively, the maximal ratio combining (MRT) is the asymptotically optimal strategy at low SNR, while ZF has good performance at high SNR or as the number of antennas increase. As discussed in [4][9], the optimal strategy lies in between these two precoders and cannot be determined without global CSI. However, inspired by the uplink-downlink duality for broadcast channels, the authors of [4] have derived a novel distributed virtual SINR precoder. The precoder vectors are achieved by maximizing the SINR-like expression in (9) where the signal power that BS b generates at UT b is balanced against the noise and interference power generated at all other UTs. It was named DVSINR as it originates from the dual virtual uplink and does not directly represent the SINR of any of the links in the downlink.

$$\mathbf{w}_{b,l}^{(DVSINR)} = \arg\max_{\|\mathbf{w}\|^2=1}\frac{\left|\mathbf{h}_{b,b,l}^H\mathbf{w}\right|^2}{\sum_{j\neq b}\left|\mathbf{h}_{b,j,l}^H\mathbf{w}\right| + \frac{\sigma^2}{P_{t_b}}} \tag{9}$$

where P_{t_b} is the per-BS power constraint. The solution to (9) is not unique, since the virtual SINR is unaffected by the phase shifts in $\mathbf{w}$. One possible solution can be written as [4],

$$\mathbf{w}_{b,l}^{(DVSINR)} = \frac{\mathbf{C}_{b,l}^{-1}\mathbf{h}_{b,b,l}}{\left\|\mathbf{C}_{b,l}^{-1}\mathbf{h}_{b,b,l}\right\|} \tag{10}$$

where

$$\mathbf{C}_{b,l}^{-1} = \frac{\sigma^2}{P_{t_b}}\mathbf{I}_{N_{t_b}} + \sum_{j\neq b}\mathbf{h}_{b,j,l}\mathbf{h}_{b,j,l}^H \tag{11}$$

IV. POWER ALLOCATION STRATEGY

In this section we design a novel distributed power allocation algorithm, based on minimization of the average BER over the available subcarriers. The criteria used to design distributed power allocation essentially lead to a redistribution of powers among subcarriers. To derive the power allocation for both precoders, we assume that the interference is negligible at both low and high SNR, even for the VSINR precoder.

The above precoders were specifically designed to make the equivalent channels, given by $h_{b,b,l}^{eq} = \mathbf{h}_{b,b,l}^H\mathbf{w}_{b,l}^{(type)}$, positive and real valued. Under free interference assumption the SINR defined in (5) reduces to,

$$\mathrm{SNR}_{b,l} = \frac{\left|\sqrt{p_{b,l}}h_{b,b,l}^{eq}\right|^2}{\sigma^2} \tag{12}$$

The above expression can be used to derive distributed power allocation because it only contains the local channel gains at BS b. Based on (6) and (12) we define the average BER as,

$$P_{av,b} = \frac{\psi}{N_c}\sum_{l=1}^{N_c}Q\left(\sqrt{\beta\,\mathrm{SNR}_{b,l}}\right) \tag{13}$$

The power allocation problem at each BS b, with per-BS power constraint, can be formulated as,

 ISBN: 978-1-61208-203-5

$$\min_{\{p_{b,l}\geq 0\}}\left(\frac{\psi}{N_c}\sum_{l=1}^{N_c}Q\left(\sqrt{\beta \mathrm{SNR}_{b,l}}\right)\right) \text{ s.t. } \left\{\sum_{l=1}^{N_c}p_{b,l}\leq P_{t_b},\forall b \quad (14)$$

The Lagrangian associated with this problem is given by,

$$L\left(p_{b,l},\mu\right)=\frac{\psi}{N_c}\sum_{l=1}^{N_c}Q\left(\sqrt{\beta \mathrm{SNR}_{b,l}}\right)+\mu\left(\sum_{l=1}^{N_c}p_{b,l}-P_{t_b}\right) \quad (15)$$

where $\mu\geq 0$ is the Lagrange multiplier [11]. Since the objective function is convex in $p_{b,l}$, and the constraint functions are linear, this is a convex optimization problem. Thus, it is necessary and sufficient to solve the Karush–Kuhn–Tucker (KKT) conditions, given by,

$$\begin{cases}\dfrac{\partial L}{\partial p_{b,l}}=\dfrac{-1}{N_c}\dfrac{\psi\beta\dfrac{h_{b,b,l}^{eq}}{\sigma}e^{-\frac{1}{2}\left(\beta\frac{h_{b,b,l}^{eq}}{\sigma}\sqrt{p_{b,l}}\right)^2}}{2\sqrt{2\pi}\sqrt{p_{b,l}}}+\mu=0\\ \dfrac{\partial L}{\partial \mu}=\sum_{l=1}^{N_c}p_{b,l}-P_{t_b}=0\end{cases} \quad (16)$$

It can be shown that the powers $p_{b,l}$ as function of the Lagrange multiplier μ are given by,

$$p_{b,l}=\frac{\sigma^2}{\beta\left(h_{b,b,l}^{eq}\right)^2}W_0\left(\frac{\psi^2\beta^2\left(h_{b,b,l}^{eq}\right)^4}{8\pi\sigma^4N_c^2\mu^2}\right) \quad (17)$$

where W_0 stands for Lambert's W function of index 0 [12]. This function $W_0(x)$ is an increasing function with $W_0(x)=0,\ x=0$ and $W_0(x)>0,\ x>0$. Therefore, μ^2 can be easily determined iteratively to satisfy $\sum_{l=1}^{N_c}p_{b,l}=P_{t_b}$, by using the bisection method. This scheme is referred as DZF *virtual* minimum BER power allocation (DZF MBER PA) or VSINR minimum BER power allocation (VSINR MBER PA) when DZF or VSINR precoders are considered, respectively.

V. NUMERICAL RESULTS

In this section, the performance of the coordinated beamforming approaches with the proposed distributed power allocation scheme will be illustrated numerically. The scenario consists of 4 uniformly distributed single antenna UTs in a square with BSs in each of the corners. The power decay is proportional to $1/r^4$, where r is the distance from a transmitter. We define the SNR at the cell edge as $\mathrm{SNR}=\mathrm{P}_{t_b}\rho_c/N_c\sigma^2$, where the ρ_c represents the long term channel power in the center of the square. This represents a scenario where terminals are moving around in the area covered by 4 base stations.

The main parameters used in the simulations are based on LTE standard [14]: FFT size of 1024; number of available subcarriers set to 128; sampling frequency set to 15.36 MHz; useful symbol duration is 66.6 μs, cyclic prefix duration is 5.21 μs; overall OFDM symbol duration is 71.86 μs; sub-carrier separation is 15 kHz, and modulation is QPSK. We used the ITU pedestrian channel model B, with the modified taps delays according to the sampling frequency defined by LTE standard.

We compare the performance results of the proposed distributed power allocation schemes, DZF MBER PA and DVSINR MBER PA. Also, these schemes are compared with equal power allocation approach, i.e., the power available at each BS is equally divided by the subcarriers, $p_{b,l}=P_{t_b}/N_c\ ,\forall(b,l)$, referred as DZF EPA and DVSINR EPA for DZF and VSINR, respectively. The results are presented in terms of the average BER as a function of Cell-edge SNR defined above.

From Fig. 2, we can see that the performance of the proposed distributed power allocation scheme, for the two precoders, outperforms their equal power i.e. the DZF EPA and DVSINR EPA approaches. This is because they redistribute the powers across the different subchannels more efficiently. As can be seen in this figure, the gains of the proposed power allocation schemes, DZF MBER PA and DVSINR MBER PA) against the equal power approaches are approximately, 8 and 6 dB (at target BER of 10^{-3}), respectively. Also, we can observe that the performance of the DZF MBER PA tends to the DVSINR MBER one as the SNR increases.

Fig. 3 shows the performance results when one more antenna is added to each BS. In this scenario the DoF of the equivalent channels variables, given by $2\left(N_{t_b}-K+1\right)$, increases from 2 (scenario one) to 4. It can be observed that

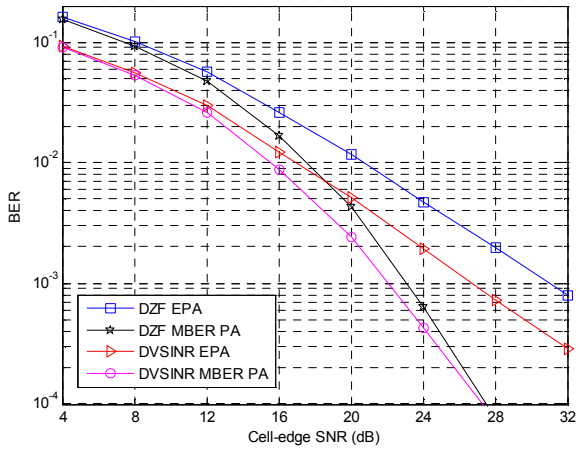


Fig. 2: Performance evaluation of the distributed precoding schemes for $N_{t_b}=4$.

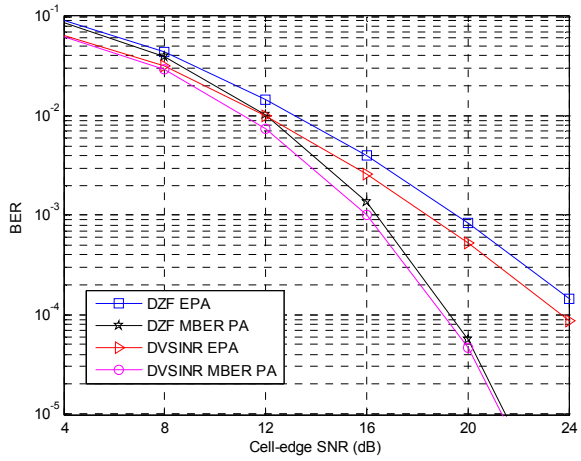


Fig. 3: Performance evaluation of the distributed precoding schemes for $N_{t_b} = 5$.

increasing the DoF, the DZF tends to the DVSINR. This behaviour is similar to the single cell systems where the precoders based on ZF criterion tends to the ones based on MMSE as the number of transmit antennas (or DoF) increases or at high SNR. From the results we can see that the gains obtained with power allocation schemes are lower, as compared with equal power approaches, than in the previous scenario.

VI. CONCLUSIONS

We proposed a novel distributed power allocation scheme for distributed precoding schemes, namely DZF and DVSINR, and for the downlink MISO-OFDM based systems. Both the precoders and power allocation were computed at each base station just by assuming the knowledge of local CSI without data sharing.

The results have shown that the proposed distributed power allocation schemes outperform the equal power ones. Also, the performance of the DZF based approaches tend to the DVSINR ones when the number of DoF increases or at high SNR.

It is clear from the presented results that the proposed distributed precoding schemes present significant interest for next generation wireless networks for which cooperation between BSs is anticipated.

ACKNOWLEDGMENT

The work presented in this paper was supported by the Portuguese CADWIN FCT project, PTDC/EEA TEL/099241/2008.

REFERENCES

[1] A. G. Armada, M. S. Fernándes, and R. Corvaja, “Waterfilling schemes for zero-forcing coordinated base station transmissions”, in *proc. of IEEE GLOBECOM*, Nov., 2009.

[2] J. Zhang, R. Chen, J. G. Andrews, A. Ghosh, and R. W. Heath Jr., ”Networked MIMO with Clustered Linear Precoding”, *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1910-1921, 2009.

[3] R. Zhang, “Cooperative multi-cell block diagonalization with per-base-station power constraints”, in *proc. of IEEE WCNC*, 2010.

[4] E. Bjornson, R. Zakhour, D. Gesbert, and B. Ottersten, “Cooperative multicell precoding: rate region characterization and distributed strategies with instantaneous statistical CSI”, *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 4298-4310, 2010.

[5] A. Silva, R. Holakouei, and A. Gameiro, “Power allocation strategies for distributed precoded multicell based systems”, *EURASIP Journal on Wireless Communications and Networking,* vol. 2011, no. 2011, April 2011.

[6] J. Lindblom, E. Karipidis, and E. G. Larsson, “Selfishness and Altruism on the MISO Interference Channel: The case of partial transmitter CSI”, *IEEE Comms. Letters*, vol. 13, no. 9, Set. 2009, pp. 667-669.

[7] J. Zhang, and J. G. Andrews, “Adaptive spatial intercell interference cancellation in multicell wireless networks”, *IEEE Journal on Selected Areas in Comms*., vol. 28, pp. 1455-1468, Dec. 2010.

[8] B. Lee, H. Je, O. S. Shin, and K. Lee, “A novel uplink MIMO transmission scheme in a multicell environment”, *IEEE Trans. On Wireless Comms.*, vol. 8, no. 10, Oct. 2009, pp. 4981-4987.

[9] R. Zakhour, D. Gesbert “Coordination on the MISO Interference Channel using the Virtual SINR Framework” International ITG/IEEE Workshop on Smart Antennas (WSA’09), Berlin, Germany, 2009.

[10] J. Proakis, *Digital Communications*, 3 rd Ed., McGrraw-Hill, New York, 1995.

[11] S. Haykin, *Adaptive Filter Theory*, 3rd Ed., Prentice Hall, 1996.

[12] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, “On the Lambert W function” *Adv. Comput. Math.,* Vol. 5, pp. 329–359, 1996.

[13] S. Boyd, and L. Vandenberghe, *Convex optimization*, Cambridge: Cambridge University Press, 2004.

[14] 3rd Generation Partnership Project, “LTE Physical Layer - General Description”, 2007, No 3. 3GPP TS 36.201 V8.1.

 ISBN: 978-1-61208-203-5

Blind Subspace Channel Estimation in MIMO-OFDM Systems with Few Received Blocks

Jung-Lang Yu, Po-Ting Chen and Wan-Ru Kuo
Department of Electrical Engineering,
Fu Jen Catholic University,
New Taipei City, 24205, Taiwan
e-mail : { yujl@ee.fju.edu.tw, kyo72929@msn.com, c022018@yahoo.com.tw }

***Abstract*—In this paper, the blind subspace channel estimation using the block matrix scheme is proposed for multiple-input multiple-output (MIMO) orthogonal frequency division multiplexing (OFDM) systems. Based on the Toeplitz structure, the block matrix scheme collects a group of the received OFDM symbols into a vector, and then partitions it into a set of equivalent signals. The number of equivalent signals is about N times of OFDM symbols, where N is the size FFT operation. The proposed blind subspace channel estimation can converge within a small amount of OFDM symbols. Besides, the semi-blind channel estimation is also examined by combining few pilot sequences with the subspace method. Simulation results show that the proposed blind and semi-blind algorithms outperform the compared methods.**

Keywords-MIMO-OFDM; blind subspace channel estimation; Toeplitz; AIC; MDL.

I. INTRODUCTION

Wideband wireless communication systems have been extensively studied in recent years for the demands of high data rate and high quality transmission. Orthogonal frequency division multiplexing (OFDM) and multiple-input multiple-output (MIMO) are two key techniques to fulfill those demands appeared in the long-term evolution (LTE) and the future fourth-generation (4G) communication systems [1]-[3]. Channel estimations are necessary for coherent detection in MIMO-OFDM systems. There are in general three categories in channel estimations which are training-based, blind and semi-blind methods, respectively. The training-based method requires extra bandwidth to accommodate the periodic known symbols and thus reduces the spectral efficiency [4][5]. The blind method saves the spectral efficiency by utilizing the statistics of received signals. But, this method requires a large amount of received signals to obtain accurate statistics [6][7]. Semi-blind methods, on the other hand, combine the blind method with few training symbols to solve the ambiguity problem occurred in blind methods [8][9].

In this paper, we discuss the blind and semi-blind subspace channel estimation for MIMO OFDM systems with much fewer received symbols. Blind subspace channel estimation has been widely examined for various precoding OFDM systems. For example, Ali et al. studied the subspace channel estimation for cyclic-prefix (CP)-OFDM, zero-padding (ZP)-OFDM and CP-free OFDM systems, respectively [10]. Li and Roy proposed subspace channel estimation based on exploiting the presence of virtual carriers for single-input single-output OFDM systems [11]. Zeng and Ng investigated in [12] the subspace channel estimation for multi-user and multi-antenna ZP-OFDM systems. Shin et al. extended the work in [11] to MIMO-OFDM systems [13]. The subspace channel estimation often converges in a large amount of received OFDM symbols. To enhance the convergence of the subspace channel estimation, Yu [14] presented the block matrix scheme (BMS) to SIMO CP-free OFDM systems. This approach can obtain a group of equivalent signals which is about N times of OFDM symbols where N is the size of FFT operation. By exploiting the idea from [14], a new block matrix scheme is applied to MIMO CP-OFDM systems, in which the number of equivalent samples is increased and the channel estimation error is lowered.

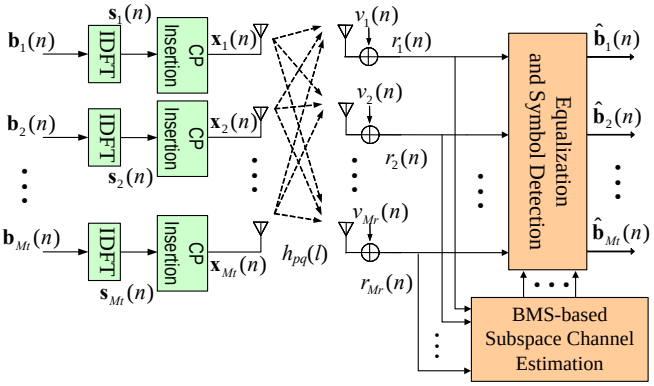


Fig.1. MIMO CP-OFDM block transmission systems. The system has M_t transmit and M_r receive antennas.

Notation: Vectors and matrices are denoted by boldface lower and upper case letters, respectively; superscripts of $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and conjugate transpose, respectively; $\mathbf{I}$ denotes an identity matrix; $\mathbf{0}$ denotes a zero vector or matrix with all zero entries; $E[\cdot]$ denotes the statistical expectation; $\|\cdot\|$ denotes the matrix or vector Frobenius norm.

The rest of this paper is organized as follows. In Section II, we introduce the signal model of the MIMO CP-OFDM systems. The subspace channel estimation is briefly described in Section III. In Section IV, the blind and semi-

 ISBN: 978-1-61208-203-5

blind subspace channel estimations are presented with the assistance of block matrix scheme. The computer simulations are performed in Section V. Finally, conclusion and future work are given in Section VI.

II. System Model

Fig. 1 shows the quasi-synchronous MIMO CP-OFDM system with M_t transmit antennas and M_r receive antennas. Let $\mathbf{b}_q(n)=\left[b_q(n,0),\cdots,b_q(n,N-1)\right]^T$ be the n-th block frequency domain symbol for the q-th transmit antenna. Transmitted symbol $b_q(n,k)$ is assumed to be independent and identically distributed (i.i.d.) complex random variable with zero-mean and variance σ_s^2 . After multicarrier modulation implemented by IDFT, the time domain signal vector is given by $\mathbf{s}_q(n)=\mathbf{W}_N^H\mathbf{b}_q(n)=\left[s_q(n,0),\cdots,s_q(n,N-1)\right]^T$ where $\mathbf{W}_N$ is the N-point DFT matrix with the (n,m)-th element $(1/\sqrt{N})\exp(-j2\pi(n-1)(m-1)/N)$. Appending CP components at the front of $\mathbf{s}_q(n)$ yields $\mathbf{x}_q(n)=\left[x_q(n,0),\cdots,x_q(n,Q-1)\right]^T$, where $Q=N+L_c$ and L_c is the length of CP. Denote by $\mathbf{x}(m)=\sum_{n=0}^{\infty}\sum_{j=0}^{Q-1}\mathbf{x}(n,j)\delta\left(m-(nQ+j)\right)$ the transmitted vector among all transmit antennas where $\mathbf{x}(n,j)=\left[x_1(n,j),\cdots,x_{M_t}(n,j)\right]^T$. The discrete-time received signal at the p-th receive antenna is given by

$$r_p(m)=\sum_{q=1}^{M_t}\sum_{l=0}^{L}h_{pq}(l)\,x_q(m-l)+v_p(m) \tag{1}$$

where $h_{pq}(l)$, l=0,…,L represents the composite channel impulse response between the q-th transmit antenna and the p-th receive antenna with maximum channel order L, and $v_p(m)$ is the additive white Gaussian noise (AWGN) with zero-mean and variance σ_n^2 . Noise is assumed to be spatially and temporally white, and be uncorrelated with transmitted symbols. In order to avoid the inter-symbol interference (ISI), we assume that $L\le L_C$. Let $\mathbf{r}(m)=\left[r_1(m),\cdots,r_{Mr}(m)\right]^T$ and stack $\mathbf{r}(m)$, $m=nQ+L_c,\cdots,nQ+Q-1$ as $\mathbf{r}_n$. Then we have

$$\mathbf{r}_n=\left[\mathbf{r}^T(nQ+L_c),\cdots,\mathbf{r}^T(nQ+Q-1)\right]^T=\mathbf{H}_N\mathbf{x}_n+\mathbf{v}_n \tag{2}$$

where $\mathbf{v}_n$ is the noise vector and

$$\mathbf{x}_n=\left[\mathbf{x}^T(n,L_c-L),\cdots,\mathbf{x}^T(n,Q-1)\right]^T\in\mathbb{C}^{(N+L)M_t\times 1}$$

$$\mathbf{H}_N=\begin{bmatrix}\mathbf{h}(L) & \cdots & \mathbf{h}(0) & \cdots & \cdots & \mathbf{0}\\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots\\ \mathbf{0} & \cdots & \cdots & \mathbf{h}(L) & \cdots & \mathbf{h}(0)\end{bmatrix} \tag{3}$$

$$\mathbf{h}(l)=\begin{bmatrix}h_{11}(l) & h_{12}(l) & \cdots & h_{1M_t}(l)\\ h_{21}(l) & h_{22}(l) & \cdots & h_{2M_t}(l)\\ \vdots & \vdots & \vdots & \vdots\\ h_{M_r1}(l) & h_{M_r2}(l) & \cdots & h_{M_rM_t}(l)\end{bmatrix}$$

Note that $\mathbf{H}_N$ is a $NM_r\times(N+L)M_t$ block Toeplitz matrix. To consider a tall and skinny matrix for $\mathbf{H}_N$, we assume that $M_r>M_t$. If it is not the case, the fractionally spaced receiver could be used here. In (2), we assume the channel order is known a priori derived from Akaike information theoretic criterion (AIC) or minimum description length (MDL) [15]. Based on the signal model in (2), we discuss the subspace channel estimation in the next section.

III. Subspace Channel Estimation

With the signal model in (2), various subspace channel estimation techniques have been presented based on different assumptions. We briefly describe the channel estimation in [12] for the purpose of comparison. Let $\mathbf{W}_N=[\mathbf{w}(0)\ \cdots\ \mathbf{w}(N\text{-}1)]$ and define $\mathbf{W}_{CP}$ and $\mathbf{W}_{CP,Mt}$ respectively by

$$\mathbf{W}_{CP}=\left[\mathbf{w}(N-L),\cdots,\mathbf{w}(N-1)\ \ \mathbf{W}_N\right]\in\mathbb{C}^{N\times(N+L)}$$

$$\mathbf{W}_{CP,Mt}=\mathbf{W}_{CP}\otimes\mathbf{I}_{Mt}\in\mathbb{C}^{NM_t\times(N+L)M_t}$$

The signal vector $\mathbf{x}_n$ can be rewritten as

$$\mathbf{x}_n=\mathbf{W}_{CP,Mt}^H\mathbf{b}_n \tag{4}$$

where

$$\mathbf{b}_n=\left[\mathbf{b}^T(n,0),\cdots,\mathbf{b}^T(n,N-1)\right]^T\in\mathbb{C}^{NM_t\times 1}$$

$$\mathbf{b}(n,j)=\left[b_1(n,j),\cdots,b_{Mt}(n,j)\right]^T\in\mathbb{C}^{M_t\times 1}$$

Substituting (4) into (2), $\mathbf{y}_n$ can be expressed by

$$\mathbf{r}_n=\mathbf{H}_N\mathbf{W}_{CP,Mt}^H\mathbf{b}_n+\mathbf{v}_n=\mathbf{H}_W\mathbf{b}_n+\mathbf{v}_n \tag{5}$$

where $\mathbf{H}_W=\mathbf{H}_N\mathbf{W}_{CP,Mt}^H$. In the subspace channel estimation, the channel is identifiable if the matrix $\mathbf{H}_W$ is of full column rank. A necessary and sufficient condition for this full column rank requirement is given in [12], which is stated as follows.

Theorem 1 [12]: In the case of $M_r>M_t$, the matrix $\mathbf{H}_W$ is of full column rank if and only if rank($\mathbf{H}(z)$)=M_t at $z=e^{j2\pi k/N}$, k=0,…,N-1, where $\mathbf{H}(z)=\sum_{n=0}^{L}\mathbf{h}(n)z^{-n}$.

From Theorem 2, we can calculate the signal and noise subspaces from $\mathbf{r}_n$ in (5) if the assumptions of $M_r>M_t$ and rank($\mathbf{H}(e^{j2\pi k/N})$)=$M_t$ are satisfied. To find the noise subspace, the correlation matrix of $\mathbf{r}_n$ is first computed by

$$\mathbf{R}_r=E[\mathbf{r}_n\mathbf{r}_n^H]=\mathbf{H}_WE[\mathbf{b}_n\mathbf{b}_n^H]\mathbf{H}_W^H+\sigma_n^2\mathbf{I}=\sigma_s^2\mathbf{H}_W\mathbf{H}_W^H+\sigma_n^2\mathbf{I} \tag{6}$$

Performing the eigenvalue-eigenvector decomposition (EVD) onto $\mathbf{R}_r$ yields the eigenvectors $\mathbf{U}$. The eigenvectors can be divided into two sets $\mathbf{U}=[\mathbf{U}_s\ \mathbf{U}_n]$ according to their eigenvalue spread, where $\mathbf{U}_s\in\mathbb{C}^{NM_r\times(N+L)M_t}$ is the signal subspace spanning the same subspace as $\mathbf{H}_W$, and $\mathbf{U}_n\in\mathbb{C}^{NM_r\times(NM_r-(N+L)M_t)}$ is the noise subspace which is

 ISBN: 978-1-61208-203-5

orthogonal to the signal subspace. Let $\mathbf{U}_n=[\mathbf{u}_1,\ldots,\mathbf{u}_{NMr-(N+L)Mt}]$ and $\mathbf{u}_k$ be partitioned into a block vector $\mathbf{u}_k=[\mathbf{u}_k^H(1)\cdots\mathbf{u}_k^H(N)]^H$ where $\mathbf{u}_k(j)$ is a $M_r\times 1$ vector. Then from the subspace orthogonal principle, we have [12]

$$\mathbf{u}_k^H\mathbf{H}_W=\mathbf{0} \text{ or } \mathbf{W}_{CP}^*\mathbf{V}_k^H\mathbf{h}=\mathbf{0} \quad (7)$$

where $\mathbf{h}=\left[\mathbf{h}^T(0),\cdots,\mathbf{h}^T(L)\right]^T$ and $\mathbf{V}_k$ is a $(L+1)M_r\times Q$ matrix

$$\mathbf{V}_k=\begin{bmatrix}\mathbf{0} & \cdots & \mathbf{0} & \mathbf{u}_k(1) & \cdots & \mathbf{u}_k(N)\\ \vdots & \iddots & \mathbf{u}_k(1) & \iddots & \mathbf{u}_k(N) & \mathbf{0}\\ \mathbf{0} & \iddots & \iddots & \iddots & \iddots & \iddots\\ \mathbf{u}_k(1) & \cdots & \mathbf{u}_k(N) & \mathbf{0} & \cdots & \mathbf{0}\end{bmatrix}$$

It is shown in [12] that the equations in (7) can determine the channel matrix **h** up to an ambiguity matrix. In practical, the correlation matrix in (6) is computed from the sample correlation matrix of $\mathbf{r}_n$. If there are K OFDM symbols available, the sample correlation matrix is given by

$$\hat{\mathbf{R}}_r=(1/K)\sum_{n=1}^{K}\mathbf{r}_n\mathbf{r}_n^H \quad (8)$$

From (8), the eigenvectors $\hat{\mathbf{u}}_k$ and matrix $\hat{\mathbf{V}}_k$ can be obtained. With the constraint that **h** has a full column rank, the channel matrix can be estimated by the least square minimization technique

$$\hat{\mathbf{h}}=\arg\min_{\mathbf{h}^H\mathbf{h}=\mathbf{I}}\sum_{k=1}^{NM_r-(N+L)M_t}\left\|\mathbf{W}_{CP}^*\hat{\mathbf{V}}_k^H\mathbf{h}\right\|_F^2 \quad (9)$$

IV. Blind Channel Estimation by Block Matrix Scheme

The estimation performance in (9) is heavily depended on the biasness of the sample correlation matrix. Let $\Delta\hat{\mathbf{R}}_r=\mathbf{R}_r-\hat{\mathbf{R}}_r$ be the bias sample correlation matrix. It is shown in [16] that the norm of the bias matrix is proportional to the dimension of $\mathbf{r}_n$, and inversely proportional to K. Therefore, the subspace channel estimation generally requires a large amount of received blocks to achieve a small perturbation of sample correlation matrix and a low channel estimation error. The block matrix scheme is proposed in this section to improve the subspace channel estimation. With the assistance of block Toeplitz structure in the received signal, the block matrix scheme segments the stacked OFDM symbols into a group of equivalent sub-vectors. The number of equivalent sub-vectors is about Q times of OFDM symbols. Therefore, the biasness of the sample correlation matrix is reduced considerably.

A. Block Matrix Scheme

We first observe that the channel matrix in (2) has a block Toeplitz form. The block matrix scheme is proposed here to increase the number of equivalent samples and then to enhance the performance of channel estimation. By collecting K consecutive received OFDM symbols, the signal vector is given by

$$\tilde{\mathbf{r}}_K=\left[\mathbf{r}^T(0),\mathbf{r}^T(1),\cdots,\mathbf{r}^T(KQ-1)\right]^T=\mathbf{H}_{QK}\tilde{\mathbf{x}}_K+\tilde{\mathbf{v}}_K \quad (10)$$

where $\tilde{\mathbf{x}}_K=[\mathbf{x}^T(-1,Q-L),\cdots\mathbf{x}^T(-1,Q-1),\mathbf{x}_0^T,\mathbf{x}_1^T,\cdots,\mathbf{x}_{K-1}^T]^T$ is a $(QKM_t+LM_t)\times 1$ vector with $\mathbf{x}_{-1}=\mathbf{0}$ and $\mathbf{H}_{QK}$ is a $QKM_r\times(QK+L)M_t$ block Toeplitz matrix which has a similar form to (3). Because of the block Toeplitz structure in $\mathbf{H}_{QK}$, we can select a proper parameter G such that $\mathbf{H}_{QK}$ is expressed by

$$\mathbf{H}_{QK}=\begin{bmatrix}\boxed{\mathbf{H}_G} & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \boxed{\mathbf{H}_G}\end{bmatrix} \quad (11)$$

where $\mathbf{H}_G$ is also a block Toeplitz matrix with dimensions $GM_r\times(G+L)M_t$. Using (10) and (11), a sub-vector $\tilde{\mathbf{r}}_{K,g}$ of $GM_r\times 1$ is defined by $\tilde{\mathbf{r}}_{K,g}=\left[\mathbf{r}^T(g),\cdots,\mathbf{r}^T(g+G-1)\right]^T$, which is obtained as

$$\tilde{\mathbf{r}}_{K,g}=\mathbf{H}_G\tilde{\mathbf{x}}_{K,g}+\tilde{\mathbf{v}}_{K,g},\quad g=0,\cdots,KQ-G \quad (12)$$

where $\tilde{\mathbf{x}}_{K,g}=\left[\mathbf{x}(g-L),\cdots,\mathbf{x}(g+G-1)\right]^T$ and $\tilde{\mathbf{v}}_{K,g}$ contains the noise components.

From the sub-vectors in (12), the subspace channel estimation can be performed if the channel matrix $\mathbf{H}_G$ has a full column rank. A tall and skinny matrix is only a necessary but not a sufficient condition for the block Toeplitz matrix to be of full column rank. That is $GM_r>(G+L)M_t$ can not guarantee that $\mathbf{H}_G$ has a full column rank. More precisely, a necessary and sufficient condition for this requirement has been presented in [17].

Theorem 2 [17]: Assume that **h**(0), **h**(L) and **H**(z) have a full column rank for all z. The block Toeplitz matrix $\mathbf{H}_G$ has a full column rank if and only if G is no less than the degree of orthogonal complement polynomial matrix of **H**(z).

If the assumptions in Theorem 2 are satisfied such that $\mathbf{H}_G$ has a full column rank, the subspace channel estimation is developed as follows. We first show that the symbols in $\tilde{\mathbf{x}}_{K,g}$ are uncorrelated. Since $\mathbf{s}_q(n)=\mathbf{W}_N^H\mathbf{b}_q(n)$, we find that $\mathbf{s}_q(n)$ is also an i.i.d. random vector because of $E[\mathbf{s}_q(n)\mathbf{s}_q^H(n)]=\mathbf{W}_N^HE[\mathbf{b}_q(n)\mathbf{b}_q^H(n)]\mathbf{W}_N=\sigma_s^2\mathbf{I}$. Denote by $\mathbf{R}_G=E[\tilde{\mathbf{r}}_{K,g}\tilde{\mathbf{r}}_{K,g}^H]$ the correlation matrix of $\tilde{\mathbf{r}}_{K,g}$. If we properly choose the parameter G such that the symbols in $\tilde{\mathbf{x}}_{K,g}$ are uncorrelated, the noise subspace can be computed from the EVD of $\mathbf{R}_G=E[\tilde{\mathbf{r}}_{K,g}\tilde{\mathbf{r}}_{K,g}^H]$

$$\mathbf{R}_G=\sigma_s^2\mathbf{H}_G\mathbf{H}_G^H+\sigma_n^2\mathbf{I}=\mathbf{E}_s\Sigma_s\mathbf{E}_s^H+\sigma_n^2\mathbf{E}_n\mathbf{E}_n^H \quad (13)$$

where $\mathbf{\Sigma}_s$ is a diagonal matrix consisting of $(G+L)M_t$ eigenvalues larger than σ_n^2, $\mathbf{E}_s$ is the signal subspace which equals the range space of $\mathbf{H}_G$, and $\mathbf{E}_n$ is the noise subspace. Using the orthogonal property between signal subspace and

noise subspace, we have $\mathbf{E}_n^H \mathbf{H}_G = \mathbf{0}$. Let $\mathbf{q}_i$ be the i-th column of $\mathbf{E}_n$ and partition $\mathbf{q}_i$ into a $G\times 1$ block vector

$$\mathbf{q}_i = \left[\mathbf{q}_i^T(0), \mathbf{q}_i^T(1), \cdots, \mathbf{q}_i^T(G-1)\right]^T \quad (14)$$

where $\mathbf{q}_i(j) \in C^{M_r}$, $j = 0, \cdots, G-1$. Exploiting the block Toeplitz structure of $\mathbf{H}_G$, $\mathbf{E}_n^H \mathbf{H}_G = \mathbf{0}$ is rewritten by

$$\mathbf{Q}_i^H \mathbf{h} = 0, \quad i = 1, \cdots, GM_r - (G+L)M_t \quad (15)$$

where $\mathbf{Q}_i \in C^{(L+1)M_r \times (G+L)}$

$$\mathbf{Q}_i = \begin{pmatrix} \mathbf{q}_i(G-1) & \cdots & \mathbf{q}_i(0) & \cdots & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & \cdots & \mathbf{q}_i(G-1) & \cdots & \mathbf{q}_i(0) \end{pmatrix} \quad (16)$$

A theorem is given in [17] that (15) can determine the channel matrix $\mathbf{h}$ up to an ambiguity matrix.

Theorem 3 [17]: Let $\mathbf{h}'$ be a matrix that has the same size as that of $\mathbf{h}$, $\mathbf{H}'_G$ be constructed from $\mathbf{h}'$ in the same form as $\mathbf{H}_G$ is constructed from $\mathbf{h}$. Assume that $\mathbf{h}(0)$, $\mathbf{h}(L)$ and $\mathbf{H}(z)$ have a full column rank for all z, and G is no less than the degree of orthogonal complement polynomial matrix of $\mathbf{H}(z)$. Then $\mathbf{h}'$ is equal to $\mathbf{h\Omega}$ where $\mathbf{\Omega}$ is an $M_t \times M_t$ invertible matrix if and only if $\mathbf{h}'$ has a full column rank and $span(\mathbf{H}'_G)$ is equal to $span(\mathbf{H}_G)$.

In a finite sample scenario, we use the sample correlation matrix $\hat{\mathbf{R}}_G$ instead of the ensemble average correlation matrix $\mathbf{R}_G$ to compute the noise subspace. Due to the biasness of the sample correlation matrix, the homogeneous equations in (15) will not be satisfied. The constrained least square optimization criterion is adopted to find the channel matrix

$$\hat{\mathbf{h}} = [\hat{\mathbf{h}}_1 \cdots \hat{\mathbf{h}}_{Mt}] = \arg \min_{\|\mathbf{h}_i\|=1} \mathbf{h}_i^H \mathbf{Q}\mathbf{Q}^H \mathbf{h}_i \quad (17)$$

where $\mathbf{Q} = \left[\mathbf{Q}_1 \cdots \mathbf{Q}_{GM_r-(G+L)M_t}\right]$. The estimates of $\mathbf{h}$ in (17) are the eigenvectors associated with the M_t smallest eigenvalues of the matrix $\mathbf{QQ}^H$. From Theorem 3, $\hat{\mathbf{h}}$ differs from $\mathbf{h}$ by an ambiguity matrix $\mathbf{\Omega}$. In the blind channel estimation, the assistance of pilot sequences is a practical way to solve the ambiguity and alleviate the phase rotation in the symbol detection.

B. Semi-blind Approach

The semi-blind estimation technique estimates the channels by combining the blind method with the pilot information [18]. From (2), $\mathbf{r}_n$ can be rewritten as

$$\mathbf{r}_n = \mathbf{H}_{cir}\mathbf{s}_n + \mathbf{v}_n$$

where $\mathbf{H}_{cir}$ is a $NM_r \times NM_t$ block circular matrix, and $\mathbf{s}_n = (\mathbf{W}_N^H \otimes \mathbf{I}_{Mt})\mathbf{b}_n$. Performing DFT operation onto $\mathbf{r}_n$ yields $\mathbf{y}_n = DFT(\mathbf{r}_n) = [\mathbf{y}^T(n,0) \cdots \mathbf{y}^T(n,N-1)]^T$ where

$$\mathbf{y}(n,k) = \mathbf{H}(k)\mathbf{b}(n,k) + \mathbf{\eta}(n,k) \quad (18)$$

, $\mathbf{H}(k) = \sum_{l=0}^{L} \mathbf{h}(l) e^{-j2\pi lk/N}$ is a $M_r \times M_t$ matrix and $\mathbf{\eta}(n,k)$ is the noise. Assume that there are A OFDM symbols and each one contains B pilots at $k_1, k_2, \ldots, k_B$ subcarriers. Define $\mathbf{Y}(n)$ and $\mathbf{B}(n)$ and $\mathbf{\eta}(n)$ respectively by

$$\begin{aligned} \mathbf{Y}(n) &= \left[\mathbf{y}(n,k_1), \mathbf{y}(n,k_2), \cdots, \mathbf{y}(n,k_B)\right]^T \\ \mathbf{B}(n) &= \left[\mathbf{b}(n,k_1), \mathbf{b}(n,k_2), \cdots, \mathbf{b}(n,k_B)\right]^T \\ \mathbf{\eta}(n) &= \left[\mathbf{\eta}(n,k_1), \mathbf{\eta}(n,k_2), \cdots, \mathbf{\eta}(n,k_B)\right]^T \end{aligned} \quad (19)$$

Then from (18) and (19), $\mathbf{Y}(n)$ is given by

$$\mathbf{Y}(n) = \sum_{l=0}^{L} \mathbf{\Phi}^l \mathbf{B}(n)\mathbf{h}^T(l) + \mathbf{\eta}(n) = \mathbf{D}(n)\tilde{\mathbf{h}} + \mathbf{\eta}(n) \quad (20)$$

where $\mathbf{\Phi} = diag(e^{-j2\pi k_1/N}, \cdots, e^{-j2\pi k_B/N})$, $\tilde{\mathbf{h}} = [\mathbf{h}(0), \cdots, \mathbf{h}(L)]^T$, $\mathbf{D}(n) = [\mathbf{B}(n), \mathbf{\Phi}^1\mathbf{B}(n), \cdots, \mathbf{\Phi}^L\mathbf{B}(n)]$. Stacking $\mathbf{Y}(n)$ for $n=n_1, n_2, \ldots, n_A$ produces

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}(n_1) \\ \vdots \\ \mathbf{Y}(n_A) \end{pmatrix} = \begin{pmatrix} \mathbf{D}(n_1) \\ \vdots \\ \mathbf{D}(n_A) \end{pmatrix} \tilde{\mathbf{h}} + \begin{pmatrix} \mathbf{\eta}(n_1) \\ \vdots \\ \mathbf{\eta}(n_A) \end{pmatrix} = \mathbf{D}\tilde{\mathbf{h}} + \mathbf{\eta} \quad (21)$$

To solve the ambiguity problem in the subspace channel estimation, we integrate linear equations in (15) and (21) together. Denote by $vec(.)$ the vectorization of a matrix by stacking its columns in order. Let $\mathbf{y}_v = vec(\mathbf{Y})$, $\tilde{\mathbf{h}}_v = vec(\tilde{\mathbf{h}})$, $\mathbf{h}_v = vec(\mathbf{h})$, and $\mathbf{\eta}_v = vec(\mathbf{\eta})$. Then (15) and (21) become

$$\begin{aligned} \mathbf{Q}_{Mt}^H \mathbf{h}_v &= 0, \\ \mathbf{y}_v &= \mathbf{D}_{Mr}\tilde{\mathbf{h}}_v + \mathbf{\eta}_v \end{aligned} \quad (22)$$

where $\mathbf{Q}_{Mt} = \mathbf{I}_{Mt} \otimes \mathbf{Q}$ and $\mathbf{D}_{Mr} = \mathbf{I}_{Mr} \otimes \mathbf{D}$. We further observe that $\tilde{\mathbf{h}}_v$ and $\mathbf{h}_v$ have the same components with different arrangement. After carefully simplification, we obtain $\tilde{\mathbf{h}}_v = \mathbf{P}_v \mathbf{h}_v$ where $\mathbf{P}_v$ is a $M_rM_t(L+1) \times M_rM_t(L+1)$ permutation matrix.

$$\mathbf{P}_v(x,y) = \begin{cases} & x = (r-1)M_t(L+1) + lM_t + t, \\ 1, & y = (t-1)M_r(L+1) + lM_r + r \\ & l = 0 \cdots L, t = 1 \cdots M_t, r = 1 \cdots M_r \\ 0, & \text{otherwise} \end{cases} \quad (23)$$

Thus the semi-blind approach can find the estimates of $\mathbf{h}$ by the following minimization criterion,

$$\hat{\mathbf{h}}_v = \arg \min_{\mathbf{h}_v} \left\|\mathbf{y}_v - \mathbf{D}_{Mr}\mathbf{P}_v\mathbf{h}_v\right\|^2 + \alpha \left\|\mathbf{Q}_{Mt}^H \mathbf{h}_v\right\|^2 \quad (24)$$

where α is a weighting constant. The solution of the problem in (24) is given by

$$\hat{\mathbf{h}}_v = (\mathbf{P}_v^H \mathbf{D}_{Mr}^H \mathbf{D}_{Mr} \mathbf{P}_v + \alpha \mathbf{Q}_{Mt} \mathbf{Q}_{Mt}^H)^{-1} \mathbf{P}_v^H \mathbf{D}_{Mr}^H \mathbf{y}_v \quad (25)$$

C. Discussions

The parameter G needs to be selected properly such that $\mathbf{H}_G$ is of full column rank. Two possible considerations are examined as follows. Firstly, the symbols in $\tilde{\mathbf{x}}_{K,g}$ are

required to be uncorrelated explained in Sec IV.A. Since $\mathbf{x}(n,i)= \mathbf{x}(n,i+N)$ for $i=0,\ldots,L_c$, those identical components will not appear in $\tilde{\mathbf{x}}_{K,g}$ at the same time if $G+L \le N$. Secondly, from Theorem 2, the necessary and sufficient condition for $\mathbf{H}_G$ has a full column rank is that G is no less than the degree of orthogonal complement polynomial matrix of $\mathbf{H}(z)$. We can show that a more practical selection to ensure that G is sufficiently large to satisfy this requirement is $G \ge M_r L$. Therefore, the selectable range of G is given by $M_r L \le G \le N-L$.

Besides, comparing with $\hat{\mathbf{R}}_r$ in (8), the sample correlation matrix for the proposed method is calculated by

$$\hat{\mathbf{R}}_G = \frac{1}{KQ-G+1} \sum_{g=0}^{KQ-G} \tilde{\mathbf{r}}_{K,g} \tilde{\mathbf{r}}_{K,g}^H \qquad (26)$$

Therefore, $\hat{\mathbf{R}}_G$ is averaged by $(KQ-G+1)$ equivalent signals with dimension of $GM_r \times 1$, while $\hat{\mathbf{R}}_r$ is averaged by K OFDM symbols with dimension of $NM_r \times 1$. According to the analysis in [16], the biasness of the sample correlation matrix for the proposed method is reduced considerably.

V. COMPUTER SIMULATIONS

Computer simulations are given here to verify the performances of the proposed channel estimations (CE). The number of subcarriers is set $N=64$ and number of CP=16. The 16QAM modulation scheme is applied. The input signal-to-noise ratio (SNR) is defined as the bit SNR at single receive antenna. The independent Rayleigh channel with exponentially decaying power delay profile of channel order $L=5$ is used in simulations. In the semi-blind approach, we use $A=2$, $B=8$ and $\alpha = 100$. The normalized root mean-squared error (NRMSE) between the estimated and true channels is given by

$$NRMSE = \sqrt{\frac{1}{N_m M_r M_t (L+1)} \sum_{p=1}^{N_m} \frac{\| \hat{\mathbf{h}}(p) - \mathbf{h}(p) \|_F^2}{\| \mathbf{h}(p) \|_F^2}} \qquad (34)$$

where the subscript p refers to the p-th simulation run and N_m denotes the number of Monte Carlo runs.

We first examine in Fig. 2 the influences of block matrix size G varied from 12 to 80 for the BMS-based channel estimators. The selectable range of G suggested in Sec IV is $10 \le G \le 59$. It is observed from Fig. 2 that the selection of G is very robust for the proposed BMS-based methods even G is larger than 59. When $G \ge 60$, there are a portion of equivalent signals suffer from the correlated transmitted signals due to CP components. However, the amount of the correlated signals is small such that it has insignificantly influence on the channel estimation. Besides, the blind method uses pilot sequences to correct the ambiguity matrix while the semi-blind method integrates the subspace information with pilot sequences in calculation of channel estimation. Therefore, the semi-blind method outperforms the blind one.

The RNMSE versus the input SNR for the compared channel estimation methods is plotted in Fig. 3. With the selection of $G=64$, the proposed methods produce almost $Q=80$ times of equivalent signals while the method in [14] has only 17 times of equivalent signals. Therefore, the proposed methods outperform the other two channel estimation methods. Furthermore, combining with 16 pilot signals for each transmit antenna, the semi-blind method obtains the lowest MSE values. Fig. 4 shows the RNMSE versus the number of OFDM symbols. As the number of OFDM symbols increases, the proposed BMS-based methods decrease the RNMSE on a steeper slope than the methods in [12] and [14]. Especially the semi-blind method converges to the lowest MSE after about $K>50$.

With the estimated channels in Figs. 3 and 4, we examine the BER performances of the minimum mean-square error (MMSE) equalizers. For the sake of comparison, the BER with real channel coefficients is also plotted. The BERs versus the input SNR are discussed in Fig. 5. Since the proposed BMS-based methods produce a lower estimation error than the compared methods, the equalizers with the former methods outperform those with the latter ones. The semi-blind method almost achieves the same BER performance as the equalizer with real channels. Finally, the BERs versus the number of OFDM symbols are examined in Fig. 6. The proposed BMS-based methods converge faster than the other two methods. Interestingly, the semi-blind method reaches the error floor of the BER at about $K=50$ in which the lowest MSE is also met.

VI. CONCLUSION AND FUTURE WORK

We proposed in this paper the blind and semi-blind subspace channel estimation for MIMO CP-OFDM systems. Inspired by the block Toeplitz structure, the block matrix scheme is first presented to increase the number of equivalent signals. The block matrix scheme decreases the biasness of the correlation matrix, noise subspace and then the channel estimation. The identifiability of the proposed channel estimation is further studied, where the estimated channels differ from the true channels by an invertible matrix. With the assistance of few pilot sequences, the semi-blind method combining the subspace method with pilot information is provided at the end. Computer simulations verify the superiority of the proposed blind and semi-blind CE over the compared ones.

We will extend the work in this paper to the OFDM systems with the virtual carriers (VCs). The VCs are often properly distributed on the dedicated band with zero values in OFDM systems for shaping the transmission spectrum and alleviating the adjacent channel interference (ACI). With the existence of VCs, the property that the symbols in $\tilde{\mathbf{x}}_{K,g}$ are uncorrelated is not hold. Additional work will be required to make the block matrix scheme applicable to OFDM systems with VCs.

 ISBN: 978-1-61208-203-5

ACKNOWLEDGMENT

This work was supported by the National Science Council, R.O.C., Taiwan under Grants NSC 99-2221-E-030-006-MY2.

REFERENCES

[1] 3GPP TS 36.201, "Long Term Evolution (LTE) physical layer; General description," Release 8, V8.3.0, March 2009.

[2] E. Dahlman, S. Parkvall, and J. Skold, *4G LTE/LTE-Advanced for Mobile Broadband*, Academic Press, 2011.

[3] M. Jiang and L. Hanzo, "Multi-user MIMO-OFDM for next generation wireless systems," *Proceedings of the IEEE*, Vol.95, No.7, pp. 1430-1469, July 2007

[4] I. Barhumi, G. Leus, and M. Moonen, "Optimal training design for MIMO OFDM systems in mobile wireless channels," *IEEE Trans. Signal Process.*, vol. 51, no. 6, pp. 1615–1624, Jun. 2003.

[5] H. Li, C. K. Ho, J. W. M. Bergmans, and F. M. J. Willems, "Pilot-aided angle-domain channel estimation techniques for MIMO-OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 57, no. 2, pp. 906–920, Mar. 2008.

[6] H. Ali, A. Doucet, and Y. Hua, "Blind SOS subspace channel estimation and equalization techniques exploiting spatial diversity in OFDM systems," *Digital Signal Processing,* vol. 14, pp. 171-202, March 2004.

[7] F. Gao, Y. Zeng, A. Nallanathan, and T. S. Ng, "Robust subspace blind channel estimation for cyclic prefixed MIMO OFDM systems: Algorithm, identifiability and performance analysis," *IEEE J. Sel. Areas Commun.*, vol. 26, pp. 378-388, Feb. 2008

[8] B. Muquet, M. de Courville and P. Duhamel, "Subspace-based blind and semi-blind channel estimation for OFDM systems," *IEEE Trans. on Signal Processing,* vol. 50, no. 7, pp. 1699–1712, Jul. 2002.

[9] F. Wan, W.-P. Zhu and M.N.S. Swamy, "Semiblind sparse channel estimation for MIMO-OFDM systems," *IEEE Trans. on Vehicular Technology,* vol. 60, no. 6, pp. 2569-2582, July 2011.

[10] H. Ali, A. Doucet, and Y. Hua, "Blind SOS subspace channel estimation and equalization techniques exploiting spatial diversity in OFDM systems," *Digital Signal Processing,* vol. 14, pp. 171-202, March 2004.

[11] C. Li and S. Roy, **"**Subspace-based blind channel estimation for OFDM by exploiting virtual carriers," *IEEE Trans. Wireless Commun.,* vol. 2, pp. 141-150, Jan. 2003.

[12] C. Shin, R. W. Heath, Jr., and E. J. Powers, "Blind channel estimation for MIMO-OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 56, no. 2, pp. 670-685, Mar. 2007.

[13] Y. Zeng and T. S. Ng, "A semi-blind channel estimation method for multiuser multiantenna OFDM systems,**"** *IEEE Trans. Signal processing,* vol. 52, pp. 1419-1429, May. 2004.

[14] J.L. Yu, "Channel Estimation for SIMO OFDM Systems without Cyclic Prefix," *Electronics Letters* vol. 43, no. 24, pp. 1369-1371, Nov. 2007

[15] G. Xu, R. H. Roy, and T. Kailath, "Detection of number of sources via exploitation of centro-symmetry property," *IEEE Trans. Signal Process.*, vol. SP-42, pp. 102-112, Jan. 1994.

[16] R. Vershynin, "How close is the sample covariance matrix to the actual covariance matrix?" [online]. Available: http://arxiv.org/PS_cache/arxiv/pdf/1004/1004.3484v1.pdf

[17] G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, *Signal Processing Advances in Wireless and Mobile Communications*, vol. 1. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[18] A. Aïssa-El-Bey, D. Kimura, H. Seki, and T. Taniguchi, "Blind and semi-blind sparse channel identification in MIMO OFDM systems," in *Proc. IEEE ICC*, Kyoto, Japan, 2011, pp. 1-5

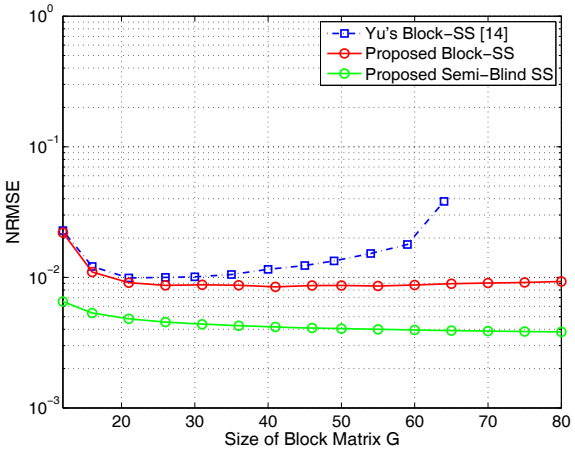


Fig. 2. NRMSE vs. Parameter G for different BMS-based channel estimation methods with SNR=20dB, K=400.

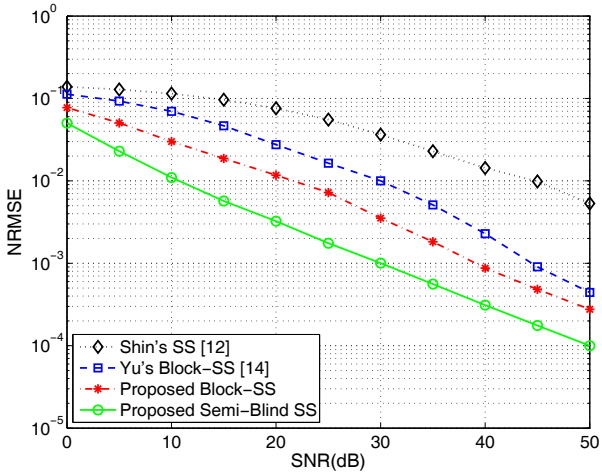


Fig. 3. NRMSE vs. SNR for compared channel estimation methods with G=64, K=200.

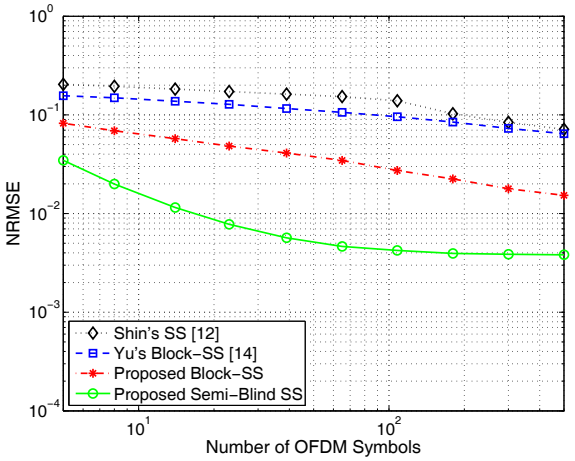


Fig. 4. NRMSE vs. the number of OFDM symbols for compared channel estimation methods with G=64, SNR=20dB.

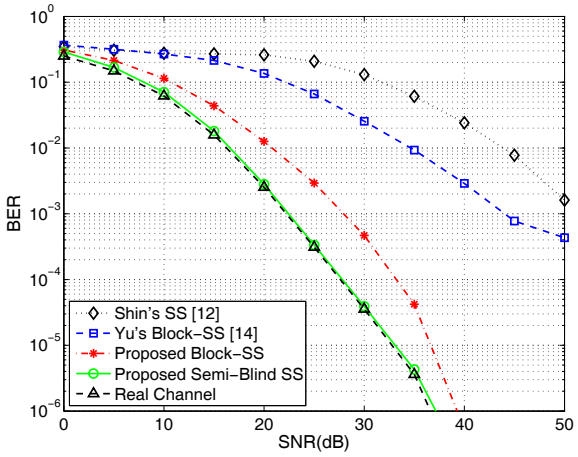


Fig. 5. BER vs. SNR for compared channel estimation methods with G=64, K=200..

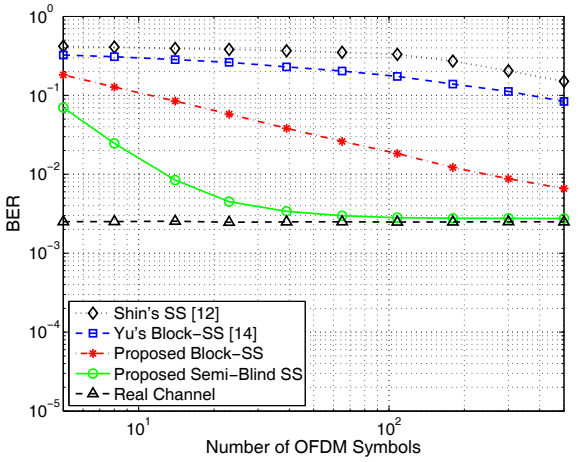


Fig. 6. BER vs. the number of OFDM symbols for compared channel estimation methods with G=64, SNR=20dB.

 ISBN: 978-1-61208-203-5

Blind Estimation Schemes for Frequency Offset of OFDM Systems in Non-Gaussian Noise Environments

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

***Abstract*—In this paper, the blind frequency offset estimation schemes robust to the non-Gaussian noise for orthogonal frequency division multiplexing (OFDM) systems are addressed. Based on the cyclic prefix structure and a maximum-likelihood (ML) estimation, the estimation schemes in non-Gaussian noise are proposed. Simulation results show that the proposed blind estimation schemes offer a robustness and a substantial performance improvement over the conventional blind estimation scheme in non-gaussian noise environments.**

Keywords-blind estimation; frequency offset; maximum-likelihood; non-Gaussian noise; OFDM.

I. INTRODUCTION

Orthogonal frequency division multiplexing (OFDM) has been adopted as a physical layer implementation in a wide variety of wireless systems such as long term evolution (LTE), wireless local area network (WLAN), and worldwide interoperability for microwave access (WiMAX) due to its immunity to multipath fading and high spectral efficiency [1]-[3]. However, OFDM is very sensitive to the frequency offset (FO) caused by Doppler shift or oscillator instabilities [1], [4]. In this paper, we focus on FO estimation based on the blind approach, which does not require an additional training symbol [4].

Assuming that the ambient noise is a Gaussian process, in [5], an optimal FO estimation scheme was proposed using the cyclic prefix (CP) of OFDM symbols without requiring the training symbol. However, it has been observed that the ambient noise often exhibits non-Gaussian nature in wireless channels, mostly due to the impulsive nature originated from various sources such as car ignitions, narrowband interferences, moving obstacles, and reflections from sea waves [6], [7]. The conventional estimation scheme developed assuming the Gaussian noise could suffer from severe performance degradation under the non-Gaussian noise environments.

In this paper, we propose robust blind FO estimation schemes in non-Gaussian noise environments. Using the CP structure of OFDM, we first derive a blind maximum-likelihood (ML) FO estimation scheme in non-gaussian noise modeled as a complex isotropic Cauchy noise, and then, propose a simpler blind estimation scheme reducing the size of the candidate set. From simulation results, the proposed schemes are confirmed to offer a substantial performance improvement over conventional blind estimation scheme in non-Gaussian noise environments.

The rest of this paper is organized as follows. Section II introduces the signal and noise model. In Section III, the novel blind FO estimation schemes are proposed, and then, in Section IV the simulation results are demonstrated. Section V concludes this paper.

II. SIGNAL MODEL

The kth received OFDM sample $r(k)$ can be expressed as

$$r(k) = x(k)e^{j2\pi k\varepsilon/N} + n(k) \tag{1}$$

for $k = -G, \cdots, -1, 0, 1, \cdots, N-1$, where $x(k)$ is the kth sample of the transmitted OFDM symbol generated by the inverse fast Fourier transform (IFFT) of size N, G is the size of the CP, ε is the FO normalized to the subcarrier spacing $1/N$, and $n(k)$ is the kth sample of additive noise.

In this paper, we adopt the complex isotropic symmetric α stable (CISαS) model for the noise samples $\{n(k)\}_{k=0}^{N-1}$ and assume that they are independent and identically distributed; this model has been widely employed due to its strong agreement with experimental data [8], [9]. For example, the interference due to the multiple access is often modeled as the SαS noise [10], [11] as well as the ambient noise in the shallow water channel of underwater communication systems [12]. The probability density function (pdf) of $n(k)$ is then given by [8]

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

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

 ISBN: 978-1-61208-203-5

A closed-form expression of (2) does not exist except for the special cases of $\alpha = 1$ (complex isotropic Cauchy) and $\alpha = 2$ (complex isotropic Gaussian). For complex isotropic Cauchy and Gaussian cases, (2) can be re-written in the closed form as

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

In this paper, we concentrate on the case of $\alpha = 1$, since it is known that the receiver which performs well in Cauchy noise also works well in other cases of SαS noise [11]. We shall see in Section IV that the estimation schemes obtained for $\alpha = 1$ are not only more robust to the variation of α, but they also provide a better performance for most values of α, than the conventional estimation scheme.

III. Proposed Schemes

A. Maximum-likelihood Blind Estimation Scheme

In estimating the FO, we consider a property of the CP structure of OFDM, i.e., $x(k) = x(k+N)$ for $k = -G, -G+1 \cdots, -1$ as in [5]. Then, from (1), we have

$$r(k+N) - r(k)e^{j2\pi\varepsilon} = n(k+N) - n(k)e^{j2\pi\varepsilon} \quad (4)$$

for $k = -G, -G+1 \cdots, -1$. Observing that $n(k+N) - n(k)e^{j2\pi\varepsilon}$ obeys the complex isotropic Cauchy distribution with dispersion 2γ (since the distribution of $-n(k)e^{j2\pi\varepsilon}$ is the same as that of $n(k)$), we obtain the pdf

$$f_{\boldsymbol{r}}(\boldsymbol{r}|\varepsilon) = \prod_{k=-G}^{-1} \frac{\gamma}{\pi\left(\left|r(k+N) - r(k)e^{j2\pi\varepsilon}\right|^2 + 4\gamma^2\right)^{\frac{3}{2}}} \quad (5)$$

of $\boldsymbol{r} = \{r(k+N) - r(k)e^{j2\pi\varepsilon}\}_{k=-G}^{-1}$ conditioned on ε. The ML estimation is then to choose $\hat{\varepsilon}$ such that

$$\begin{aligned} \hat{\varepsilon} &= \arg\max_{\tilde{\varepsilon}} [\log f_{\boldsymbol{r}}(\boldsymbol{r}|\tilde{\varepsilon})] \\ &= \arg\min_{\tilde{\varepsilon}} \Lambda(\tilde{\varepsilon}), \end{aligned} \quad (6)$$

where $\tilde{\varepsilon}$ denotes the candidate value of ε and the log-likelihood function $\Lambda(\tilde{\varepsilon}) = \sum_{k=-G}^{-1} \log\left\{\left|r(k+N) - r(k)e^{j2\pi\tilde{\varepsilon}}\right|^2 + 4\gamma^2\right\}$ is a periodic function of $\tilde{\varepsilon}$ with period 1: The minima of $\Lambda(\tilde{\varepsilon})$ occur at a distance of 1 from each other, causing an ambiguity in estimation. Assuming that ε is distributed equally over positive and negative sides around zero, the valid estimation range of the ML estimation scheme can be set to $-0.5 < \varepsilon \leq 0.5$, as in [5]. The estimation scheme (6) will be called the Cauchy ML blind estimation (CMBE) scheme.

B. Low-complexity Blind Estimation Scheme

The CMBE scheme is based on the exhaustive search over the whole estimation range ($|\varepsilon| < 0.5$), which requires high computational complexity. Thus, we propose a low-complexity FO estimation scheme with the reduced set of the candidate values.

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

$$\bar{\varepsilon}(k) = \frac{1}{2\pi}\angle\{r^*(k)r(k+N)\}, \text{ for } k = -G, -G+1 \cdots, -1. \quad (7)$$

Exploiting the set of the candidate values in (7), the FO estimate $\hat{\varepsilon}_L$ can be obtained as follows

$$\hat{\varepsilon}_L = \arg\min_{\bar{\varepsilon}(k)} \Lambda\big(\bar{\varepsilon}(k)\big), \text{ for } k = -G, -G+1 \cdots, -1. \quad (8)$$

In the following, (8) is denoted as the low-complexity CMBE (L-CMBE) scheme. Using only $N/2$ candidate values, the L-CMBE scheme can offer an almost same performance as the CMBE scheme with the exhaustive search, which is verified by simulation results in Section IV.

IV. Simulation Results

In this section, the proposed CMBE and L-CMBE schemes are compared with the Gaussian ML blind estimation (GMBE) scheme in [5] in terms of the mean squared error (MSE). We assume the following parameters: The IFFT size $N = 64$, FO $\varepsilon = 0.25$, the search spacing of 0.001 for the CMBE scheme, and a multipath Rayleigh fading channel with length $L = 8$ and an exponential power delay profile of $\mathbf{E}[|h(l)|^2] = \exp(-l/L)/\{\sum_{l=0}^{L-1}\exp(-l/L)\}$ for $l = 0, 1, \cdots, 7$, where $h(l)$ is the lth channel coefficient of a multipath channel and $\mathbf{E}[\cdot]$ denotes the statistical expectation. Since CISαS noise with $\alpha < 2$ has an infinite variance, the standard signal-to-noise ratio (SNR) becomes meaningless for such a noise. Thus, we employ the geometric SNR (GSNR) defined as $\mathbf{E}[|x(k)|^2]/(4C^{-1+2/\alpha}\gamma^{2/\alpha})$, where $C = \exp\{\lim_{m\to\infty}(\sum_{i=1}^{m}\frac{1}{i} - \ln m)\} \simeq 1.78$ is the exponential of the Euler constant [13]. The GSNR indicates the relative strength between the information-bearing signal and the CISαS noise with $\alpha < 2$. Clearly, the GSNR becomes the standard SNR when $\alpha = 2$. Since γ can be easily and exactly estimated using only the sample mean and variance of the received samples [14], it may be regarded as a known value: Thus, γ is set to 1 without loss of generality.

Figs. 1-3 show the MSE performances of the CMBE, L-CMBE, and GMBE schemes as a function of α when the GSNR is 5, 10, and 15 dB, respectively. From the figures, we can clearly observe that the proposed schemes not only outperform the conventional scheme for most values of α, except for those close to 2, but also provide a robustness

 ISBN: 978-1-61208-203-5

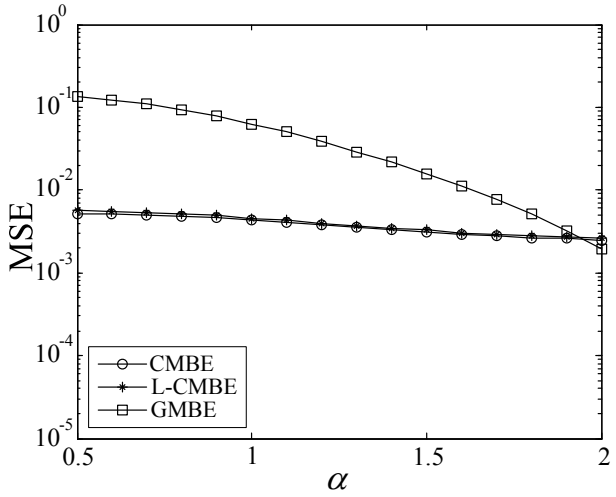


Figure 1. The MSE performances of the CMBE, L-CMBE, and GMBE schemes as a function of α when the GSNR is 5 dB.

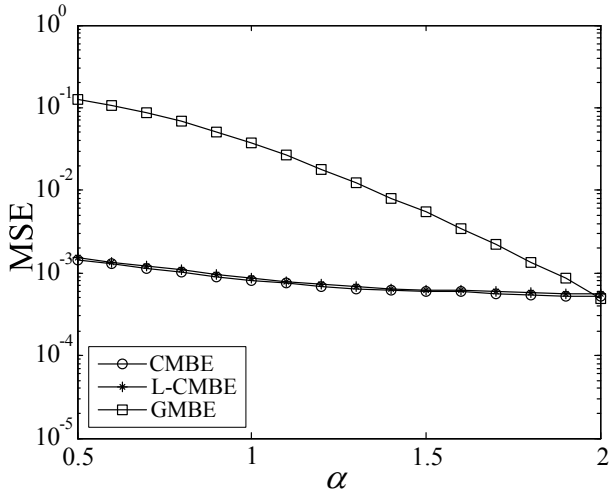


Figure 2. The MSE performances of the CMBE, L-CMBE, and GMBE schemes as a function of α when the GSNR is 10 dB.

to the variation of the value of α. Another important observation is that the estimation performance of the L-CMBE scheme is almost same as that of the CMBE scheme. From this observation, it is confirmed that the candidate values for the L-CMBE scheme is reasonable.

Fig. 4 shows the MSE performances of the proposed and conventional schemes as a function of GSNR when $\alpha = 1$. From the figure, we can clearly observe that the proposed schemes outperform the conventional scheme regardless of value of GSNR. Moreover, the MSE performance of L-CMBE is very similar to that of the CMBE, the optimal FO estimation scheme when $\alpha = 1$, but also provide a robustness to the variation of the value of α.

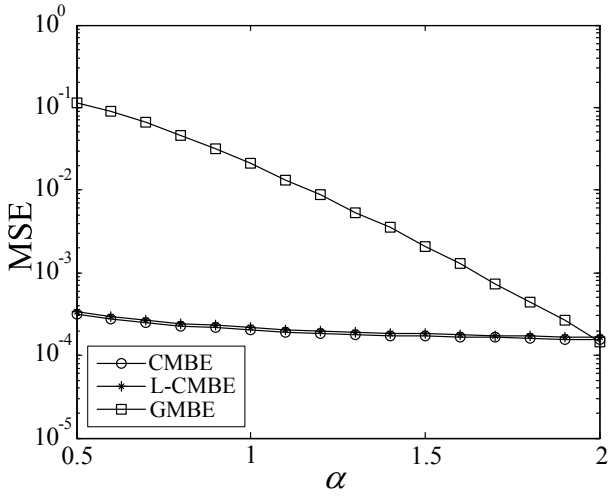


Figure 3. The MSE performances of the CMBE, L-CMBE, and GMBE schemes as a function of α when the GSNR is 15 dB.

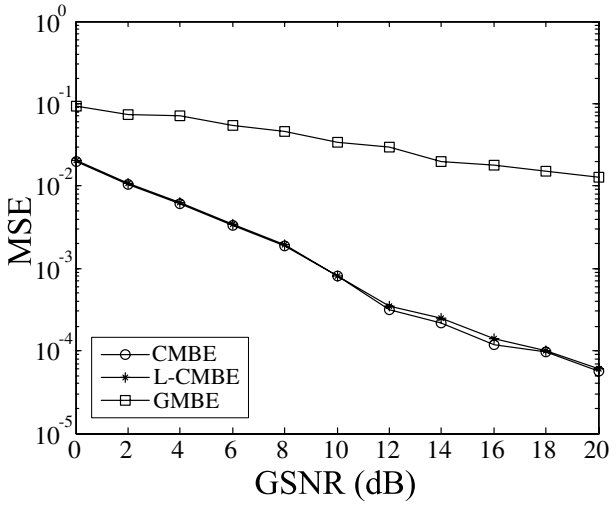


Figure 4. The MSE performances of the CMBE, L-CMBE, and GMBE schemes as a function of GSNR when $\alpha = 1$.

V. Conclusion

In this paper, we have proposed blind FO estimation schemes in non-Gaussian noise environments. Based on the CP structure of OFDM, we first have derived a ML FO estimation scheme in non-gaussian noise modeled as a complex isotropic Cauchy noise, and then, derived a simpler blind estimation scheme with a lower complexity. From simulation results, it has been confirmed that the proposed schemes offer a robustness and a substantial performance improvement over the conventional estimation scheme in non-gaussian noise environments.

Acknowledgment

This research was supported by the National Research Foundation (NRF) of Korea under Grant 2011-0018046

 ISBN: 978-1-61208-203-5

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

REFERENCES

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

[2] Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specification: spectrum and transmit power management extensions in the 5GHz band in Europe, IEEE, 802.11h, 2003.

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

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

[5] J.-J. Beek, M. Sandell, and P. O. Borjesson, "ML estimation of time and frequency offset in OFDM systems," *IEEE Trans. Sig. Process.,*, vol. 45, no. 7, pp. 1800-1805, July 1997.

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

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

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

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

[10] J. Ilow, D. Hatzinakos, and A. N. Vanetsanopoulos, "Performance of FH SS radio networks with interference modeled as a mixture of Gaussian and alpha-stable noise," *IEEE Trans. Commun.*, vol. 46, no. 4, pp. 509-520, Apr. 1998.

[11] S. Kalyani and K. Giridhar, "OFDM channel estimation in the presence of NBI and the effect of misspecified NBI model," in *Proc. Workshop on Signal Process. Wireless Commun. (SPAWC)*, pp. 1-5, Helsinki, Finland, June 2007.

[12] M. Chitre, S. H. Ong, and J. Potter, "Performance of coded OFDM in very shallow water channels and snapping shirimp noise," in *Proc. MTS/IEEE Oceans*, pp. 1-6, Washington, DC, Sep. 2005.

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

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

 ISBN: 978-1-61208-203-5

A Periodogram-based CFO Estimation Scheme for OFDM Systems

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

***Abstract*—In this paper, we propose a novel carrier frequency offset (CFO) estimation scheme for orthogonal frequency division multiplexing (OFDM) systems in non-Gaussian noise environments. The proposed scheme has much wider estimation range compared with that of the conventional scheme, improving the overall CFO estimation performance. Numerical results demonstrate that the proposed scheme has better estimation performance than the conventional scheme.**

Keywords-estimation; carrier frequency offset; orthogonal frequency division multiplexing (OFDM); non-Gaussian; periodogram.

I. INTRODUCTION

Tho orthogonal frequency division multiplexing (OFDM) signal has been widely used for wireless communication systems including wireless local area network (WLAN), wireless metropolitan area network (WMAN), and digital video broadcasting (DVB) systems [1]. However, the OFDM system is very sensitive to the carrier frequency offset (CFO) caused by Doppler shift or oscillator instabilities. Thus, the CFO estimation is one of the most important issues in OFDM systems.

Several schemes [2]-[4] have been proposed to estimate the CFO of OFDM signals. Schmidl and Cox proposed a CFO estimation scheme using a training symbol with two identical halves [2], whose estimation range is equal to the sub-carrier spacing. In [3], a new CFO estimation scheme that utilizes a training symbol with more than two identical parts was proposed, increasing the estimation range twice that of the scheme in [2]. With the maximum-likelihood (ML) criterion, in [4], an optimal scheme for CFO estimation was derived using the same training symbol as in [3]. Recently, in [5], a periodogram-based CFO estimation scheme was proposed, whose estimation range is as large as the bandwidth of the OFDM signal while maintaining the same level of the estimation performance as those of the schemes based on training symbols with identical parts. However, the conventional schemes are developed under the assumption of the Gaussian distributed noise. Since it has been observed that the noise often exhibits non-Gaussian nature in wireless channels [6], the conventional estimators could suffer from performance degradation in the non-Gaussian noise environments.

In this paper, we propose a novel periodogram-based CFO estimation scheme for OFDM systems in non-Gaussian noise environments. We first investigate the influence of the non-Gaussian noise on the integer part of CFO estimation scheme in [5], and then, propose a novel fractional frequency offset (FFO) estimation scheme with wider estimation range. The numerical results show that the proposed FFO estimation scheme has better estimation performance than the FFO estimation scheme in [5] under the influence of the non-Gaussian noise.

II. SIGNAL MODEL

After the inverse fast Fourier transform (IFFT) operation, at the transmitter, the nth complex-valued OFDM sample $x(n)$ can be expressed as

$$x(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k e^{j2\pi kn/N}, \quad n = 0, 1, \cdots, N-1, \quad (1)$$

where N is the size of the IFFT and X_k is a phase shift keying (PSK) or a quadrature amplitude modulation (QAM) symbol in the kth sub-carrier. The data part of the OFDM symbol has a duration of T seconds, and the cyclic prefix (CP), whose length is generally designed to be longer than the channel impulse response, is inserted to avoid the intersymbol interference (ISI).

The nth received OFDM sample $r(n)$ is obtained by sampling the received OFDM signal every $T_s = T/N$ seconds and can be expressed as

$$r(n) = s(n)e^{j2\pi(\varepsilon_I+\varepsilon_F)n/N} + w(n), \quad (2)$$

where $s(n) = \sum_{k=0}^{L-1} h_k x(n-k)$ is the signal component with the kth channel filter tap coefficient h_k and the channel memory size L, ε_I and ε_F represent the integer FO (IFO) and FFO normalized to the sub-carrier spacing $1/T$, respectively, and $w(n)$ is the non-Gaussian noise sample. In this paper, we assume the static channel during one OFDM symbol duration and perfect timing synchronization.

 ISBN: 978-1-61208-203-5

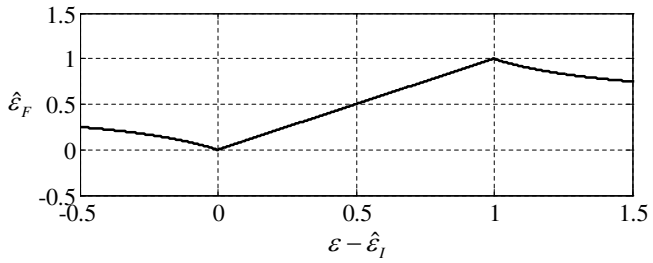


Figure 1. $\hat{\varepsilon}_F$ as a function of $\varepsilon - \hat{\varepsilon}_I$ in [5].

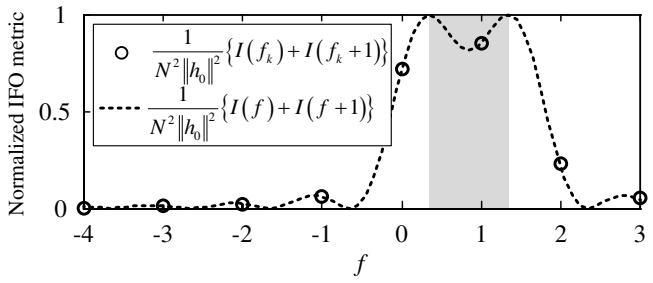


Figure 2. IFO metric $\{I(f) + I(f+1)\}$ normalized to $N^2\|h_0\|^2$ as a function of the frequency $f \in [-N/2, N/2)$ for $\varepsilon_F = 0.3$ when $\varepsilon_I = 1$, $N = 8$, and the noise is absent.

III. Proposed Scheme

A. Influence of the non-Gaussian noise on the IFO estimation

In [5], the estimates $\hat{\varepsilon}_I$ and $\hat{\varepsilon}_F$ of the IFO and FFO are obtained as

$$\hat{\varepsilon}_I = \arg\max_{f_k} \big\{I(f_k) + I(f_k+1)\big\} \tag{3}$$

and

$$\hat{\varepsilon}_F = \frac{\sqrt{I(\hat{\varepsilon}_I+1)}}{\sqrt{I(\hat{\varepsilon}_I)} + \sqrt{I(\hat{\varepsilon}_I+1)}}, \tag{4}$$

respectively, where 'arg' is the argument operation and $I(f_k)$ is the signal periodogram defined as

$$I(f_k) = \left| \sum_{n=0}^{N-1} r(n)e(n)e^{-j2\pi f_k n/N} \right|^2, \tag{5}$$

where $f_k \in \{-\frac{N}{2}, -\frac{N}{2}+1, \cdots, \frac{N}{2}-1\}$ is the kth IFO candidate, $|\cdot|$ is the absolute operation, and $e(n)$ is the envelope equalized processing factor, which removes the influence of the data modulation, defined by $\frac{x(n)^*}{\|x(n)\|^2}$ with the complex conjugation '$*$' and Euclidean norm $\|\cdot\|$.

In the absence of the noise, $\hat{\varepsilon}_F$ is given by $\frac{Z(\hat{\varepsilon}_I)}{Z(\hat{\varepsilon}_I)+Z(\hat{\varepsilon}_I+1)}$, where $Z(\alpha) = |\sin(\pi(\varepsilon-\alpha)/N)|$, and is drawn as a function of $\varepsilon - \hat{\varepsilon}_I$ as shown in Fig. 1, where $\varepsilon = \varepsilon_I + \varepsilon_F$. It is seen from the figure that the FFO can be correctly estimated only when $0 \le \varepsilon - \hat{\varepsilon}_I < 1$ (i.e., $\hat{\varepsilon}_I \in (\varepsilon - 1, \varepsilon]$) which is referred to as the correct estimation range.

Fig. 2 shows the IFO metric $\{I(f) + I(f+1)\}$ normalized to $N^2\|h_0\|^2$ as a function of the frequency $f \in [-N/2, N/2)$ for $\varepsilon_F = 0.3$ when $\varepsilon_I = 1$, $N = 8$, and the noise is absent, where '$\circ$' represents the IFO metric value corresponding to each f_k and the shaded region represents the correct estimation range. In this paper, the correct estimation probability of the IFO is defined as the probability that the maximum IFO metric corresponds to f_k stays within the correct estimation range.

From the figure, we can see that the IFO metric value (outside the correct estimation range) nearest to the correct estimation range is as large as that in the correct estimation range. Thus, under the influence of non-Gaussian noise with impulsive nature, the IFO estimation scheme (3) often outputs an incorrect estimate.

B. Proposed FFO Estimation Scheme

First, we define a new function similar to the periodogram, which can be expressed as

$$I_p(f_k) = \sum_{n=0}^{N-1} r(n)e(n)e^{-j2\pi f_k n/N}. \tag{6}$$

In the absence of the noise and interferences, $I_p(f_k)$ can be re-written as

$$I_p(f_k) = \sum_{n=0}^{N-1} h_0 e^{j2\pi(\varepsilon - f_k)n/N}. \tag{7}$$

Next, using the ratio $I_p(\hat{\varepsilon}_I)$ to $I_p(\hat{\varepsilon}_I+1)$, we can remove the channel component h_0 as follows

$$\begin{aligned}\frac{I_p(\hat{\varepsilon}_I)}{I_p(\hat{\varepsilon}_I+1)} &= \frac{h_0 \cdot \frac{1-e^{j2\pi(\varepsilon-\hat{\varepsilon}_I)}}{1-e^{j2\pi(\varepsilon-\hat{\varepsilon}_I)/N}}}{h_0 \cdot \frac{1-e^{j2\pi(\varepsilon-\hat{\varepsilon}_I)}e^{-j2\pi}}{1-e^{j2\pi(\varepsilon-\hat{\varepsilon}_I)/N}e^{-j2\pi/N}}} \\ &= \frac{1-e^{j2\pi(\varepsilon-\hat{\varepsilon}_I)/N}e^{-j2\pi/N}}{1-e^{j2\pi(\varepsilon-\hat{\varepsilon}_I)/N}}.\end{aligned} \tag{8}$$

From (8), the estimate of the FFO $\hat{\varepsilon}_F$ ($\approxeq \varepsilon - \hat{\varepsilon}_I$) can be obtained as

$$\hat{\varepsilon}_F = \frac{N}{2\pi}\angle\left(\frac{1-M(\hat{\varepsilon}_I)}{e^{j2\pi/N} - M(\hat{\varepsilon}_I)}\right), \tag{9}$$

where $M(\hat{\varepsilon}_I) = I_p(\hat{\varepsilon}_I)/I_p(\hat{\varepsilon}_I+1)$ and $\angle(y)$ denotes an angle of y. From (9), we can see that the estimation range of the proposed scheme is $-\frac{N}{2} \le \hat{\varepsilon}_F < \frac{N}{2}$. Thus, the proposed scheme (9) can estimate the FFO in both correct IFO estimate and incorrect IFO estimate cases.

IV. Numerical Results

In this section, we compare the performance of the proposed and conventional [5] FFO estimation schemes in terms of the correct estimation probability. In the simulation, we assume the following parameters: quadrature PSK (QPSK) data modulation, the FFT size of $N = 64$, a CP with a length of 8 samples, and the maximum Doppler shift of 125 Hz (corresponding to a mobile speed of 54 km/h with a carrier

 ISBN: 978-1-61208-203-5

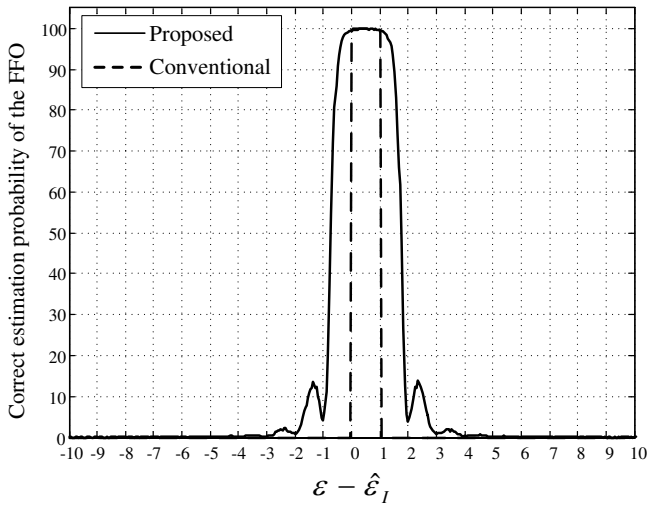


Figure 3. Correct estimation probabilities of the FFO as a function of $\varepsilon - \hat{\varepsilon}_I$ for the proposed and conventional schemes in the Rayleigh fading channel model when G-SNR is 5 dB.

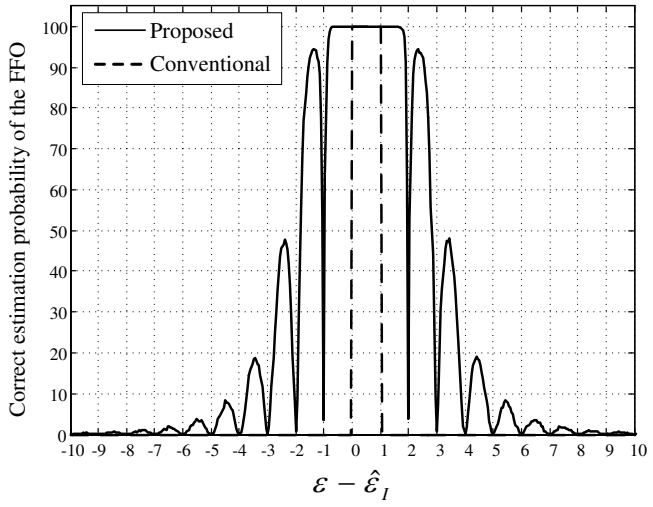


Figure 4. Correct estimation probabilities of the FFO as a function of $\varepsilon - \hat{\varepsilon}_I$ for the proposed and conventional schemes in the Rayleigh fading channel model when G-SNR is 25 dB.

frequency of 2.5 GHz). The non-Gaussian noise is modeled as Cauchy noise [7]. Since the variance is not defined in Cauchy noise, the standard signal-to-noise ratio (SNR) becomes meaningless. Thus, we employ the geometric SNR (G-SNR), which provides a mathematically and conceptually valid characterization of the relative strength between the signal and noise [8], is defined as

$$\text{G-SNR} = \frac{\sigma_s^2}{2C\gamma^2}, \tag{10}$$

where $\sigma_s^2 \triangleq \mathbf{E}\{|s(n)|^2\}$ with the expectation operator $\mathbf{E}\{\cdot\}$, $C = \exp\{\lim_{b\to\infty}(\sum_{a=1}^{b}\frac{1}{a} - \ln b)\} \approxeq 1.78$ is the exponential of the Euler constant, and γ is the dispersion parameter, which is set to be 1 in this paper. We consider four-path Rayleigh fading channel model with path delays of 0, 2, 4, and 6 samples and exponential power delay profile of $\mathbf{E}\{A_l^2\} = \exp(-0.768l)$ for $l = 0, 1, 2$, and 3 (i.e., the power ratio of the first and last paths is set to 10 dB).

Figs. 3 and 4 show the correct estimation probabilities of the FFO as a function of $\varepsilon - \hat{\varepsilon}_I$ for the proposed and conventional schemes. In the figures, we can see that the proposed scheme has wider estimation range than the conventional scheme. Thus, the proposed scheme can estimate the FFO when $\varepsilon - \hat{\varepsilon}_I < 0$ and $\varepsilon - \hat{\varepsilon}_I > 1$ (outside of the correct estimation range of the IFO). However, we can observe that the estimation range of the proposed scheme is narrower than that in the ideal case of $-\frac{N}{2} \leq \hat{\varepsilon}_F < \frac{N}{2}$ in the absence of noise. This can be explained as follows. The proposed scheme is based on $I_p(\hat{\varepsilon}_I)$ and $I_p(\hat{\varepsilon}_I + 1)$. As shown in Fig. 2, when the chosen $\hat{\varepsilon}_I$ is far from the correct estimation range of the IFO, $I_p(\hat{\varepsilon}_I)$ and $I_p(\hat{\varepsilon}_I + 1)$ have small values, and thus, they would suffer from more severe influence of the non-Gaussian noise. Thus, the correct estimation probability of the proposed scheme becomes smaller as $|\varepsilon - \hat{\varepsilon}_I|$ increases and the estimation range of the proposed scheme is narrower than that in the ideal case. However, as mentioned in Section III-A and shown in Fig. 2, in the most cases, the proposed scheme can improve the overall CFO estimation performance.

In passing, we would like to stress that the proposed scheme is applicable to any kind of OFDM system employing training symbol. However, in the applications adopting the training symbol solely dedicated to the frequency offset estimation (e.g., IEEE 802.11 [9], long term evolution (LTE) [10]), the performance of the proposed scheme may not be better than the schemes dedicated to the application only.

V. Conclusion

In this paper, we have proposed a novel CFO estimation scheme for OFDM systems in the non-Gaussian noise environments. We have first investigated the influence of the non-Gaussian noise on the IFO estimation, and then, proposed a novel FFO estimation scheme with wider FFO estimation range. From numerical results, we have confirmed that the proposed scheme has better estimation performance than the conventional scheme.

Acknowledgment

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

REFERENCES

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

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

[3] S. Chang and E. J. Powers, "Efficient frequency-offset estimation in OFDM-based WLAN systems," *Electron. Lett.*, vol. 39, no. 21, pp. 1554-1555, Oct. 2003.

[4] M.-H. Cheng and C.-C. Chou, "Maximum-likelihood estimation of frequency and time offsets in OFDM systems with multiple sets of identical data," *IEEE Trans. Sig. Process.*, vol. 54, no. 7, pp. 2848-2852, July 2006.

[5] G. Ren, Y. Chang, H. Zhang, and H. Zhang, "An efficient frequency offset estimation method with a large range for wireless OFDM systems," *IEEE Trans. Vehic. Technol.*, vol. 56, no. 4, pp. 1892-1895, July 2007.

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

[7] P. Tsakalides and C. L. Nikias, "Maximum likelihood localization of sources in noise modeled as a stable process," *IEEE Trans. Sig. Process.*, vol. 43, no. 11, pp. 2700-2713, Nov. 1995.

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

[9] Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specification: spectrum and transmit power management extensions in the 5GHz band in Europe, IEEE, 802.11h, 2003.

[10] S. Caban, C. Mehlfuhrer, M. Rupp, and M. Wrulich, *Evaluation of HSDPA and LTE: From Testbed Measurements to System Level Performance*. John Wiley and Sons, 2011.

 ISBN: 978-1-61208-203-5

A Novel Concept of UWB Pulse Switching in Sensor Networks

Qiong Huo and Subir Biswas
Michigan State University, USA
qionghmsu@gmail.com, sbiswas@egr.msu.edu

***Abstract* -This paper presents the initial results of applying a novel concept of energy-efficient pulse switching protocol for ultra-light-weight wireless network applications. The key idea is to abstract a single pulse, as opposed to multi-bit packets, as the information exchange mechanism. Pulse switching is shown to be sufficient for event sensing applications with binary sensing. Event sensing with conventional packet transport can be prohibitively energy-inefficient due to the communication, processing, and buffering overheads of the large number of bits within a packet's data, header, and preambles. The paper presents the key architectural ideas of a joint MAC-Routing protocol for pulse switching with a novel hop-angular event localization.**

Keywords-Impulse Radio; Pulse Switching; UWB; Sensor Network; Event Monitoring; Pulse Routing

I. INTRODUCTION

The key idea in this paper is to introduce a new abstraction of *pulse switching* for replacing the traditional packet switching for event monitoring. An example application is *Structural Health Monitoring* (SHM) [1] in which while monitoring a bridge for structural failures, it may be sufficient for a sensor to generate an event to indicate a structural crack in its vicinity. Sending an event, indicating the presence of the crack, to a sink would require single bit information transport. For this scenario, packets can be energy inefficient due to the communication, processing, and buffering overheads of a large number of bits within the payload, header, and the synchronization preambles [2] in each packet.

In the proposed pulse switching paradigm, such an event can be coded as a single pulse, which is then transported multi-hop while preserving the event's localization information. The resulting operational lightness, leveraged via zero collision, zero buffering, no addressing, no packet processing, and ultra-low communication and energy budgets makes the protocol applicable for severely resource-constrained sensor devices such as Radio Frequency Identifiers (RFIDs) operating with tight energy budgets, often from harvested energy [3].

The primary challenges are: 1) how to transport localization information using a single pulse, and 2) how to route a pulse without being able to explicitly code any information within the pulse. A key architectural novelty in this work is to integrate a pulses' (i.e. event's) location of origin within the MAC-routing protocol syntaxes. More specifically, by observing the time of arrival of a pulse with respect to the MAC-routing frame, a sink can resolve the corresponding event location with a pre-set resolution. Multi-hop pulse routing is addressed by introducing the concept of a novel *synchronized phase waves* across different hop-distances from the sink.

The rest of the paper is organized as follows. Section *II* presents the related work. Section *III* describes the network, application, and localization model. Section *IV* presents the MAC-Routing pulse switching architecture, and Section *V* presents simulated performance results of the proposed architecture. Finally, Section *VI* summarizes the paper.

II. RELATED WORK

Jain et al. [4] reduce preamble and header overheads by aggregating payloads from multiple *short* packets into a single *large* packet that is routed to a sink node. While reducing the energy cost, aggregation still requires the inherent packet overheads. The objective of our work is to fully eliminate packets via replacing them by pulse switching.

Fragouli et al. [5] develop models for energy and delay bounds for bit (i.e., packet based) and pulse communications in single hop networks. The main results are to demonstrate that the worst case energy performance of pulse communication can be substantially better than that of packet based communication, although with a possibly worse delay performance. A notable limitation of this work is that the paper does not provide mechanisms for scaling these results for multi-hop networks. Also, no Medium Access Control (MAC) and routing protocol details are provided for pulse switching. The objective of this paper is to design a MAC-routing framework that can be used for practical implementations of multi-hop pulse switching.

III. NETWORK AND APPLICATION MODEL

Network Model: As shown in Fig. 1, a network contains arbitrarily distributed sensors that send pulses to a sink. Depending on the node locations and the transmission range (assumed to be non-uniform), each node resides at a certain *hop-distance* from the sink. In Fig. 1, the hop-distance for each node is marked under the node. The sink is assumed to be capable of making high-power transmissions with full network coverage for frame-synchronizing the sensors.

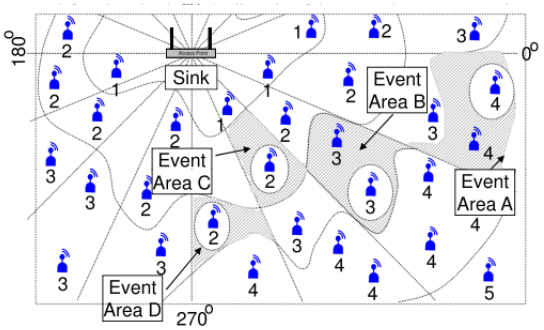


Figure 1. Network model with hop-angular localization

Application Model: Pulse switching can be used for event sensing applications. An event results in multiple pulses, which are transported multi-hop to a sink. A pulse is able to represent: a) the very occurrence of the event, and b) its location of origin. With this information, several high level conclusions can be derived at the sink by correlating multiple event pulses.

Hop-angular Event Localization: A concept of *hop-angular event area* is introduced for event localization. The network is logically divided into a fixed number of angular sectors. In Fig. 1, for example, there are 16 22.5°-wide sectors. With a pre-

 ISBN: 978-1-61208-203-5

defined sector-width (α^o), the location of a sensor can be represented by the tuple {*sector-id, hop-distance*}. For example, the location of the encircled sensor in Event Area *B* in Fig. 1 can be represented as {*15, 3*}, meaning the node is located in the 15th sector, with a hop-distance 3 from the sink.

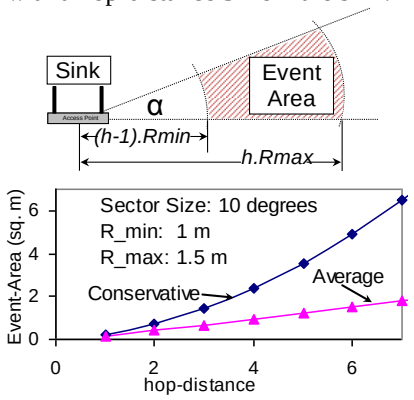


Figure 2. Hop-distance event localization

The concept of event area does not assume any specific shape (i.e. circular or otherwise) of a node's transmission coverage area. It could be of any arbitrary shape as shown in Fig. 1. While the angle for a node is pre-programmed at the deployment time, its hop-distance can be dynamically discovered using the process outlined in Section IV.C. The {*sector-id, hop-distance*} tuple indicates an *event-area*, whose size determines the event localization resolution. This tuple for an event's origin is carried to the sink by the corresponding pulse.

Consider the example event-area identified by the tuple {α, *h*} in the top portion of Fig. 2. With a sector-width of α, and R_{min}, R_{max} representing the known minimum and maximum wireless transmission range, the most conservative (coarse) localization resolution can be expressed as the largest possible event area: $A_{conservative} = \{h^2 R_{\max}{}^2 - (h-1)^2 R_{\min}{}^2\}\alpha\pi/360$. The average resolution is $A_{averge} = \{h^2 R^2 - (h-1)^2 R^2\}\alpha\pi/360$, where $R = (R_{\min} + R_{\max})/2$. For example, with transmission range spanning between 1m to 1.5m, in a network with sector-width (i.e. α) of 10°, the size of an event-area that is 5 hop-distances away is approximately 3.5 square meters. For the Structural Health Monitoring application on a bridge, this means that a structural crack can be localized within approximately 3.5 square meter area. For a given α and transmission range, since this resolution reduces with higher hop-distances, the maximum network size will be determined based on the desired resolution.

IV. Pulse Switching Architecture

A. Joint MAC-Routing Frame Structure

Nodes are frame-by-frame time synchronized by the sink, and they maintain joint MAC-Routing frames (see Fig. 3) in which each slot is used for sending a single pulse. The slot includes a guard-time to accommodate the cumulative clock-drift during a frame, which can be very small for RF technology such as Ultra Wide Band (UWB) Impulse Radio, as the frame size itself can be ultra-short (μs) for UWB. As shown in Fig. 3, the frame contains an uplink part and a downlink part. The uplink part contains a control sub-frame and an event sub-frame. The downlink part of the frame contains a synchronization slot in which the sink transmits a *full power* pulse to make all nodes frame-synchronized. The two following downlink slots and the reconfiguration part of the uplink control sub-frame are used for hop-distance discovery. The reconfiguration area in control sub-frame has (*H+1*) slots, where *H* is the maximum hop-distance. The forwarding flag area is designed for routing pulses towards the sink. The *H*-slot wide routing area of the control sub-frame is used for energy management. These functions will be explained in detail in the next few sections.

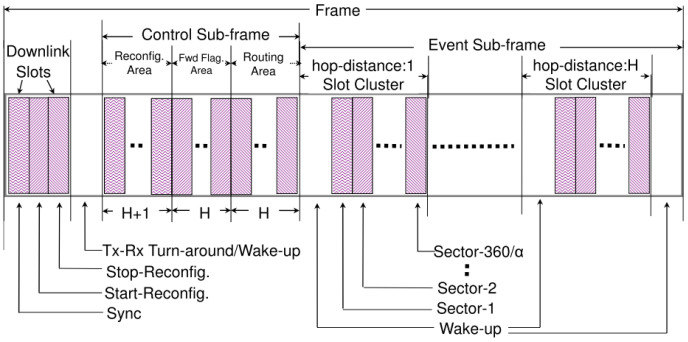


Figure 3. MAC-Routing frame for multi-hop pulse switching

The event sub-frame contains *H* slot clusters, each cluster containing *360/α* slots, where α corresponds to the sector-width. Each slot within a cluster corresponds to a specific {*sector-id, hop-distance*} tuple. Meaning, for each event-area, represented by {*sector-id, hop-distance*}, there is a dedicated slot in the event sub-frame. An event originating node transmits a pulse during the dedicated event sub-frame slot that corresponds to the {*sector-id, hop-distance*} of the node's event area. While routing the pulse towards the sink, at all intermediate nodes it is transmitted at the same event sub-frame slot that corresponds to the {*sector-id, hop-distance*} of its event-area of origin. In other words, while being forwarded, the transmission slot for the pulse at all intermediate nodes does not change with respect to the frame. This is how information about the location of origin of an event is preserved during routing. Upon reception, the sink can infer the event-area of origin from the {*sector-id, hop-distance*} value corresponding to the slot at which the pulse is received. The role of the control sub-frame in Fig. 3 will be described in Sections IV.C, IV.D and V.

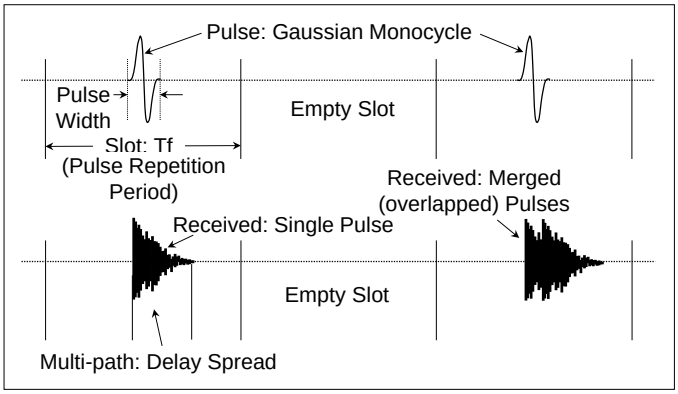


Figure 4. Pulse switching with un-modulated UWB pulses

B. Pulse Realization using UWB Impulse Radio

The ability to transmit and receive a single pulse without per-pulse synchronization overhead is a key requirement for pulse switching. Ultra Wide Band (UWB) [6][7] Impulse Radio (IR)

technology can be practically [8] used because of its support of single pulse transmission and reception. The top graph in Fig. 4 depicts a UWB implementation of the slot structure used in this protocol. A typical UWB pulse width is *1 ns*, and the pulse repetition period T_b is *1000 ns* [6], which determines the slot size in this case.

C. Pulse Forwarding

A hop-distance discovery process needs to be periodically executed by the network for each node to discover its own hop-distance from the sink node. When a pulse is transmitted by a node at hop-distance *h*, only its neighboring nodes at hop-area (*h-1*) forward it towards the sink. Meaning, the nodes at hop-area *h* and *h+1* should ignore the pulse. This logic ensures that a pulse is eventually delivered to the sink. While transmitting a pulse in the event sub-frame (see Section IV.A), its transmitter also sends a pulse in the corresponding slot of the *forwarding flag* area of the control sub-frame. That is, while forwarding a pulse by a hop area *h* node, it sends a pulse in the h^{th} time slot of the *forwarding flag* area. By looking at the received pulse in the *forwarding area*, all the receivers of the pulse can decide if it should be discarded or forwarded towards the sink. This can ensure that a pulse from hop-area *h* should be forwarded only by nodes in hop-area *h-1*.

D. Sector-constrained Routing

The extent of *sector-constraints* during pulse forwarding can be parameterized using δ, which represents the ratio of the angular resolution α and an angle γ. The angle γ is the sector-width beyond which a pulse may not be flooded while forwarding. For a given α, the minimum and the maximum values of γ are α and 180^o respectively.

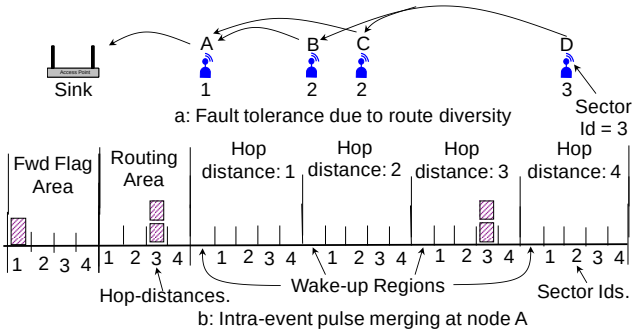


Figure 5. Route diversity and pulse merging/aggregation

The corresponding δ values are 1 and $180^o / \alpha$. When δ is 1, routing is maximally constrained, indicating the minimum communication energy consumption, and the maximum susceptibility to errors due to the minimum pulse redundancy, as shown in Fig. 6.

E. Aggregation via Pulse Merging

Collisions between pulses may not necessarily lead to information loss. For example, in Fig. 5, since the pulse originating from D is transmitted by *B* and *C* on the same slot in the event sub-frame, the receiver *A* detects RF signals for a merged pulse in that slot. As long as the RF hardware can detect the presence of this overlapped pulse, the routing continues. In fact, this pulse merging and route diversity provides inherent in-network aggregation for events from the same event-area.

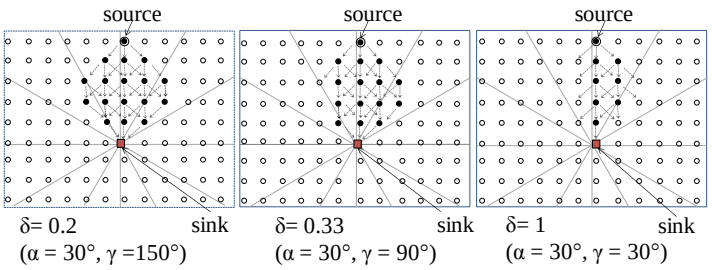


Figure 6. Routing envelope and pulse fan-out during forwarding

V. Performance Evaluation

We developed an event-driven C++ simulator which implements MAC framing and pulse routing using the UWB Impulse Radio model as presented in Section IV.

A. Pulse Transmission Count

For the proposed Pulse Switching Protocol (PSP), Fig. 7:a reports the number of forwarding transmissions across different hop areas when an event is created in hop area 5 in the 441 node network. Numbers are reported for two different sector constraints (δ=0.2 and 1). For both δ observe how the number of pulse transmissions maximizes at an intermediate hop-distance, confirming the lateral fan-out and convergence seen as the routing envelopes in Fig. 6.

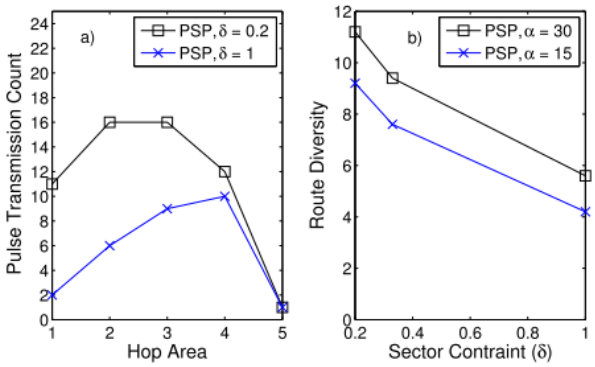


Figure 7. Transmission count and route diversity

B. Route Diversity

The route diversity factor β ($\beta \geq 1$) represents the number of forwarding transmissions for a pulse from hop-distance *h*, normalized by *h*, which is the minimum number of required transmissions. Fig. 7b demonstrates β with increasing sector-constraint δ for two different angular resolutions of α = 15° and 30°. A larger α means more nodes are involved in forwarding (see Fig. 6), leading to higher route diversity.

C. Error Analysis

Impacts of Pulse Loss Error: Pulse losses can manifest in the form of un-reported events. We define *Pulse Loss Rate (PLR)* as the probability that a pulse is lost in a given time-slot due to multi-path, channel noise, or interferences. *Event Loss Rate (ELR)* is the probability that a pulse is not reported to the sink. Fig. 8:a depicts simulation results of PLR versus ELR for a single event generated in hop area 5 of the 441-node network. Observe that for practical range of PLR [10], the ELR for PSP remains vanishingly small and it is generally insensitive to the value of PLR. This is mainly because of the redundancy in pulse transmissions (i.e. route diversity) for PSP as demon-

strated in Fig. 7.

Impacts of False Positive Errors: If pulses are erroneously detected [9] by a node in state *LO* or *TL* such that a false positive pulse in the control sub-frame corresponds to another false positive pulse in the event sub-frame with corresponding forwarding flag (see Fig. 3), then an event is falsely detected at that node. Once such a false positive event is generated, it is forwarded all the way to the sink, leading to a false positive event reporting. Let *FPPR* (False Positive Pulse Rate) be the probability that a false positive pulse is generated due to faulty UWB detection in a given time-slot. We intend to determine the quantity *FPEGR* (False Positive Event Generation Rate) which corresponds to the probability of at least one false positive event generation per frame per node at a given hop-area.

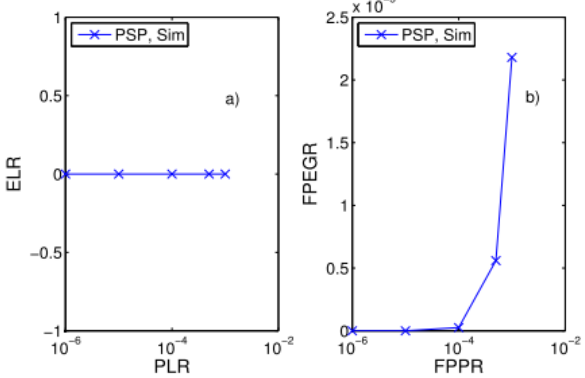


Figure 8. Impacts of pulse loss and false positive

The impacts of *FPPR* on *FPEGR* in hop area 3 in PSP are shown in Fig. 8:b. Hop area 3 is chosen because it represents the middle of the experimental network. Observe that in PSP, *FPEGR* is extremely small with practical range of *FPPR* [10] which is lower than 10^{-4}. This indicates that the proposed PSP is fairly immune to false positive errors.

D. Energy Consumption

A distributed TDMA with sink-rooted minimum spanning tree for packet routing has been used as the representative protocol to compare its energy consumption with that of the proposed pulse switching. TDMA is chosen because of its high energy efficiency compared to random access mechanisms. An event detected by a sensor is reported to sink using min-hop routing along the minimum spanning tree. A packet contains the minimum amount of information to represent a {*sector-id, hop-distance*} event-area and also a per-packet preamble [2].

Based on the UWB specification [8], the transmission and reception consumptions are set to *4 nJ* and *8 nJ* per pulse. Since a pulse transmission using the baseband UWB has the same energy expenditure for Pulse Position Modulated bits in a packet, the same *4 nJ* and *8 nJ* values are used for both pulse and bit (in packets) transmission and reception.

Fig. 9 reports the communication power consumption for both pulse and packet transport with varying event rates. Observe that the consumption is linearly dependent on the event rate λ for both pulse and packet scenarios. The slope of the TDMA graph in Fig. 9 is noticeably higher than that for PSP for both angular resolutions of $\alpha = 15^{\circ}$ and 30°, and it is mainly due to the overhead of per-packet preamble and payload overheads. Overall, Fig. 9 validates the primary premise of pulse switching that it can transport multi-point-to-point binary events at a lower energy budget compared to traditional packet switching.

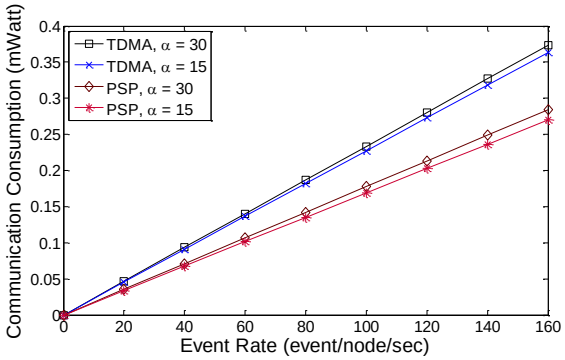


Figure 9. Communication power consumption per node

VI. Conclusions and Future Work

A novel pulse switching protocol for ultra-light-weight networking applications has been developed in this paper. A joint MAC-routing architecture for pulse switching with a hop-angular event localization strategy was presented. Through simulation results, it is shown that the proposed pulse switching architecture can be an effective means for energy efficiently transporting information that is binary in nature.

Future work on this topic includes: 1) an implementation of the proposed pulse routing architecture on a UWB embedded hardware, 2) extending the architecture for cellular event localization, and 3) experimenting with pulse switching for non-radio media such as ultrasound on metal substrates.

VII. Acknowledgement

This work was supported by an NSF grant NeTS 0915851.

VIII. References

[1] C.R. Farrar, G. Park, D.W. Allen, and M.D. Todd, "Sensor network paradigms for structural health monitoring", J. of Structural Control and Health Monitoring, vol. 13, no. 1, 2006, pp. 210-225.

[2] Z. Yuanjin, C. Rui, L. Yong, "A new synchronization algorithm for UWB impulse radio communication systems", Int. Conf. on Communication Systems (ICCS), Singapore, 2004, pp. 25-29.

[3] E. Minazara, D. Vasic, and G. Poulin, "Piezoelectric diaphragm for vibration energy harvesting", Ultrasonics, vol. 44, no. 1, 2006, pp. 699-703.

[4] A. Jain, M. Gruteser, and D. Grunwald, "Benefits of Packet Aggregation in Ad-Hoc Networks", Technical Report CU-CS-960-03, Dept. of Computer Science, Boulder, Colorado, August, 2003.

[5] C. Fragouli and A. Orlitsky, "Silence is golden and time is money: power-aware communication for sensor networks", in Allerton Conference on Comm., Control and Computing, 2005.

[6] S. Haykin and M. Moher, "Modern wireless communications", Prentice-Hall, Inc., 2004, Upper Saddle River, NJ, USA.

[7] M. Win and R. Scholtz, "Impulse radio: how it works", IEEE Communication Letters, vol. 2, no. 2, 1998, pp. 36-38.

[8] B. Poucke and B. Gyselinckx, "Ultra-wideband communication for low-power wireless body area networks", Industrial Embedded Systems Resources Guide, 2005.

[9] Van Trees and Harry L., "Detection, Estimation, and Modulation Theory - Part l." John Wiley & Sons.

[10] Yu. Andreyev, A. Dmitriev, E. Efremova, A. Khilinsky and L. Kuzmin, "Qualitative Theory of Dynamical Systems, Chaos and Contemporary Wireless Communications", International Journal of Bifurcation and Chaos, Vol.15, No.11, 2005, pp. 3639-3651.

 ISBN: 978-1-61208-203-5

Efficient Interpolation Architecture for Soft-Decision List Decoding of Reed-Solomon Codes

Sungman Lee
The MTH
426-5, Gasan-dong, Kumcheon-gu
Seoul, Korea
e-mail: orozi318@naver.com

Taegeun Park
The Catholic University of Korea
San43-1, Yokok2-dong, Bucheon-shi,
Kyungki-do, Korea
e-mail: parktg@catholic.ac.kr

***Abstract*— Recently, algebraic soft-decision decoding algorithm for RS codes that can correct the errors beyond the error correcting bound has been proposed. The main task in the algorithm is the weighted interpolation of a bivariate polynomial that requires intensive computations. In this paper, we propose an efficient architecture with low hardware complexity for interpolation in soft-decision list decoding of Reed-Solomon codes. The proposed architecture processes the candidate polynomial in such a way that the terms of *X* degrees are processed in serial and the terms of *Y* degrees are processed in parallel. The processing order of candidate polynomials adaptively changes to increase the efficiency of memory access for coefficients. The proposed interpolation architecture for the (255, 239) RS list decoder is designed and synthesized using the DonbuAnam 0.18um standard cell library. The maximum operating clock frequency is 200MHz and the synthesized gate count is about 25.1K gates in two-input equivalent NAND gates.**

Keywords—VLSI architecture; Polynomial interpolation; Reed-Solomon codes; Soft-decision list decoding.

I. INTRODUCTION

Among the various kinds of error correcting codes in digital communication systems, Reed-Solomon (RS) codes are widely used block codes due to their excellent burst error-correcting capabilities. It is well known that an (n, k) RS codes have k message symbols and n coded symbols, where each symbol belongs to $GF(2^m)$. An (n, k) RS codes can correct ν symbols and ρ erasures with $2\nu + \rho \leq n - k$. Classical RS decoding scheme can be thought of as the bounded minimum distance (BMD) algorithm that decodes the received codewords through the channel by hard-decision. Efficient algebraic hard-decision decoding algorithms, such as Berlekamp-Massey algorithm and Euclid algorithm [1] have been widely used to decode the Reed-Solomon codes.

Recently, Guruswami-Sudan (GS) [2] proposed a polynomial-time list decoding algorithm that can correct the errors beyond the error correcting bound. The proposed list decoding algorithm has a decoding radius $t' > \lfloor d_{min}/2 \rfloor$ and corrects up to $n - \sqrt{nk}$ errors for all code rates [2]. With reliable soft-decision data, such as probabilistic channel information, RS decoding can achieve better performance in correcting errors. Koetter and Vardy (KV) [3] generalized the list decoding algorithm that can decode, as long as a certain weighted condition is satisfied. The soft-decision list decoding algorithm consists of two major processes: interpolation process with KV front end and factorization process. The interpolation process is quite computationally intensive with large latency, so it may suffer low performance. The re-encoding scheme can be applied to reduce the number of iterations for interpolation [4].

Many researchers have proposed a number of the interpolation architectures for the soft-decision RS decoder [5][6][7][13][14]. Most of the architectures proposed so far try to increase the decoding performance by parallelizing the processes for the candidate polynomials. This requires considerable hardware with memory modules that may be hard to apply in some applications. Ahmed, Koetter, and Shanbhag proposed the point-serial algorithm that calculates all the discrepancy coefficients corresponding to a particular interpolation point in parallel [5]. Wang and Ma represent the finite field numbers in both regular and power formats, i.e., hybrid-format, that can reduce the hardware complexity for the DCC block and parallelize the decoder architecture [6]. The parallel architecture [7] proposed by Gross, Kschischang, and Gulak embeds both a binary tree and a linear array in a 2–D array processor, enabling fast polynomial evaluation operations. Zhu *et. al.* have proposed backward interpolation, which eliminates interpolation points or reduces interpolation multiplicities [13]. The proposed architectures share computational units with forward interpolation architectures to reduce the hardware complexity. In [14], new techniques are employed to achieve high-speed interpolation for the iterative bit-level generalized minimum distance (BGMD) algebraic soft-decision decoding. They also proposed architectures to efficiently integrate the combined and backward interpolation techniques.

In this paper, we propose an efficient architecture with low hardware complexity for interpolation in a soft-decision list decoding of Reed-Solomon codes. To reduce hardware cost, the proposed architecture processes the terms of X degrees in the candidate polynomial serially, whereas it processes the terms of Y degrees in the candidate

This work was partially supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2008-313-D00743)

This work was partially supported by the Research Fund, 2011 of The Catholic University of Korea.

 ISBN: 978-1-61208-203-5

polynomial in parallel. During the polynomial update in the interpolation process, the appropriate polynomial coefficients should be read efficiently from memory. The processing order of candidate polynomials adaptively changes to increase the efficiency of memory access for coefficients. This scheduling minimizes the usage of internal registers and the number of memory accesses and simplifies the memory structure by combining and storing data in memory. Also, the proposed architecture shows high hardware efficiency, since each module is balanced in terms of latency and the modules are maximally overlapped in the schedule.

II. SOFT-DECISION LIST DECODING OF RS CODES

An (n,k) RS codes defined in the Galois field $GF(2^m)$ have codewords of length $n = 2^m - 1$, where m is a positive integer and k is the number of information symbols in the codeword. RS codes can be obtained by evaluating certain subspaces of $\mathbb{F}_q(X)$ in a set of points $\boldsymbol{\mathcal{P}} = \{x_0, x_1, \cdots, x_{n-1} \subseteq \mathbb{F}_q(X)\}$. Therefore, (n,k) RS codes $\mathbb{C}_q(n,k)$ of length n and dimension k is defined as

$$\mathbb{C}_q(n,k) = \{(f(x_0), f(x_1), \cdots, f(x_{n-1})) : x_0, x_1, \cdots, x_{n-1} \subseteq \boldsymbol{\mathcal{P}}, f(x) \in \mathbb{F}_{\boldsymbol{q}}(\boldsymbol{X}), \deg f(X) < k\} \quad (1)$$

Guruswami and Sudan verified that the generalized Reed-Solomon decoding problem reduces to the polynomial reconstruction problem [2]. Now, we define the essential elements of the soft decision list decoding algorithm. The bivariate polynomial $Q(X,Y)$ over $\mathbb{F}_q$ is defined as in [3].

$$Q(X,Y) = \sum_t^r \sum_s^{w_v} q_{s,t} X^s Y^t = \theta_0(X) + \theta_1(X)Y + \theta_2(X)Y^2 + \cdots + \theta_v(X)Y^v + \cdots + \theta_r(X)Y^r,$$

where $\theta_v(X) = q_{0,v} + q_{1,v}X + q_{2,v}X^2 + \cdots + q_{w_v,v}X^{w_v}$, $0 \le v \le r$. (2)

We define the weighted degree of a polynomial, as follows.

Definition 1: Let $Q(X,Y) = \sum_i^\infty \sum_j^\infty q_{i,j} X^i Y^j$ be a bivariate polynomial over $GF(2^q)$, and let w_x, w_y be real numbers. Then, the (w_x, w_y)-weighted degree of $Q(X,Y)$, denoted $\deg_{w_x,w_y} Q(X,Y)$, is the maximum over all numbers $iw_x + jw_y$, such that $q_{i,j} \neq 0$.

Definition 2: A bivariate polynomial $Q(X,Y)$ is said to pass through a point (x_i, y_i) with multiplicity m_{x_i,y_i}, if the shifted polynomial $Q(X - x_i, Y - y_i)$ contains a monomial of degree m_{x_i,y_i} and does not contain a monomial of degree less than m_{x_i,y_i}. Equivalently, the point (x_i, y_i) is said to be a zero of multiplicity m_{x_i,y_i} of the polynomial $Q(X,Y)$.

When $P(X,Y)$ is the shifted version of the polynomial $Q(X,Y)$ by (x_i, y_i), the equation below holds.

$$P(X,Y) = \sum_\beta^r \sum_\alpha^{w_v} p_{\alpha,\beta} X^\alpha Y^\beta = Q(X - x_i, Y - y_i) = \sum_\beta^r \sum_\alpha^{w_v} q_{\alpha,\beta} (X - x_i)^\alpha (Y - y_i)^\beta.$$

$$coef(Q(X - x_i, Y - y_i), X^\alpha Y^\beta) = p_{\alpha,\beta} \quad , \quad p_{\alpha,\beta} = 0, \forall\, \alpha + \beta < m. \quad (3)$$

The soft-decision RS list decoding can be considered as a "curve-fitting" problem. In the first phase, the algorithm finds a polynomial $Q(x,y)$ of low degree that fits the points (x_i, y_i). Next, it finds all small degree roots of $Q(x,y)$; and each factor of $Q(x,y)$ forms possible candidates of the message polynomial.

We now briefly describe the list decoding algorithm. Figure 1 shows the block diagram of the soft-decision RS list decoder. The block diagram consists of three steps: multiplicity computation, interpolation, and factorization. The multiplicity computation step calculates the multiplicity matrix that has reliability information from the channel.

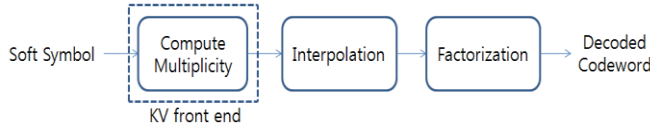


Figure 1. Block diagram of soft-decision RS list decoder.

Let us suppose that we have the set of interpolation points $(x_i, y_{i,j}), 1 \le i \le 2^q, 1 \le j \le n$ with corresponding multiplicities $m_{i,j}$. The interpolation step forms the nontrivial polynomial $Q(x,y)$ of minimal $(1, k-1)$-weighted degree that passes each interpolation point $(x_i, y_{i,j})$ with multiplicity at least $m_{i,j}$. This bivariate polynomial $Q(x,y)$ may contain the message polynomial as a root. After the bivariate polynomial $Q(x,y)$ is found, the factorization step determines all the factors of $Q(x,y)$ in the form of $Y - f(X)$, where the degree of $f(X)$ is at most k. Each root polynomial $f(X)$ is the candidate of the message polynomial. Finally, the polynomial with the highest probability among the candidates is selected as a message polynomial.

Interpolation algorithm

- Initialization
  $Q_v(X,Y) = Y^v$, *for* $0 \le v \le r$.
- Iteration for each point set (x_i, y_i, m_{x_i,y_i})
  $O_v = (w_x, w_y) - \text{weighted degree of } Q_v(X,Y)$, for $0 \le v \le r$.
  for $\beta = 0$ to $m_{x_i,y_i} - 1$
    for $\alpha = 0$ to $m_{x_i,y_i} - 1 - \beta$
      - DCC(Discrepancy Coefficient Computation)
        $d_v^{(\alpha,\beta)} = coef(Q_v(X + x, Y + y), X^\alpha Y^\beta)$, for $0 \le v \le r$
      - If there exist $\eta = \arg\min_{0 \le v \le r,\ d_v^{(\alpha,\beta)} \neq 0} \{O_v\}$
      - PU(Polynomial Update)
        $Q_v(X,Y) = d_\eta^{(\alpha,\beta)} Q_v(X,Y) + d_v^{(\alpha,\beta)} Q_\eta(X,Y)$, for $v \neq \eta$
        $Q_v(X,Y) = Q_v(X,Y)(X + x)$, for $v = \eta$
        $O_\eta = O_\eta + 1$
    end for
  end for

Figure 2. Interpolation algorithm.

III. PROPOSED ARCHITECTURE

Now, we will explain the interpolation algorithm in more detail. The interpolation step finds the bivariate polynomial that fits the set of points with the corresponding multiplicities. Two main interpolation algorithms have been proposed so far: a constrained-serial interpolation algorithm [6][7][8] and a point-serial interpolation algorithm [5]. The point-serial interpolation algorithm is usually less efficient and less flexible in architecture than the constraint-serial interpolation algorithm [6]. Figure 2 shows the interpolation

 ISBN: 978-1-61208-203-5

algorithm based on the Fundamental Iterative Algorithm (FIA) [10].

The algorithm consists of two major operations, namely the Discrepancy Coefficient Computation (DCC) and the Polynomial Update (PU). As we explained earlier, the interpolation process finds a bivariate polynomial $Q(x,y)$ that passes a point (x_i, y_j) with a multiplicity $m_{i,j}$. In the algorithm, the DCC operation computes the discrepancy coefficients corresponding to a particular constraint for each candidate polynomial and the PU operation updates the polynomial by reducing the corresponding discrepancy coefficients to zero.

A. Proposed interpolation architecture and its scheduling

Figure 3 shows the block diagram of the proposed interpolation architecture. The proposed architecture consists of the Discrepancy Coefficient Computation Unit (DCCU) that calculates the discrepancy coefficients (DC) using the Hasse Derivative (HD), the Polynomial Update Unit (PUU) that updates the candidate polynomials, the Polynomial Order Sorting Unit (POSU) that decides the processing order of data by storing and aligning the weighted degrees of the candidate polynomials, and a few memory blocks and control logic. The DCCU also consists of the HD computation block and the Y-generator that calculates the power of Y for the HD computation. The DCCU takes the stream of interpolation points and multiplicities as inputs.

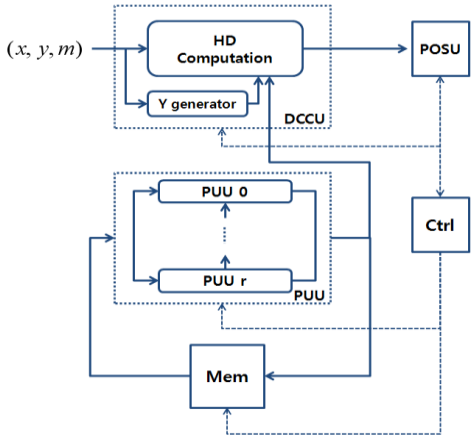


Figure 3. Proposed interpolation architecture.

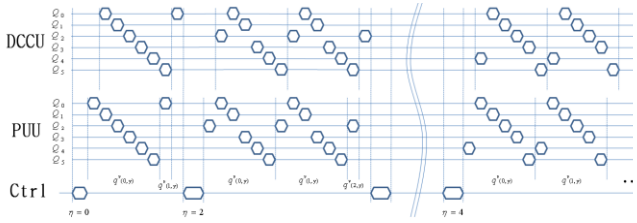


Figure 4. Timing diagram showing overlap between the DCCU and the PUU when $r = 5$.

The proposed architecture parallelizes the computation for the terms in Y and computes each candidate polynomial and the monomials in X sequentially and thus the required hardware is reduced. When the polynomials are updated sequentially, the "pivot" polynomial $Q_\eta(X,Y)$, which has the smallest weighted degree and a non-zero DC value, is used to update the other polynomials and is updated itself last.

Figure 4 shows the processing schedule of DCCU and PUU when $r = 5$, where Q_v denotes a candidate polynomial and $q^v_{(i,j)}$ denotes the coefficient of X^iY^j in Q_v. All the coefficients of X, with the same degree in Q_v, $(q^v_{(x,0)}, q^v_{(x,1)}, q^v_{(x,2)}, \dots, q^v_{(x,r)})$, are processed simultaneously, since the proposed architecture processes the terms in Y in parallel. We can overlap the computation of the DCCU and the PUU by sending the output coefficient of the PUU directly to the DCCU, as depicted in Fig.4. The updated coefficients in candidate polynomials are stored and sent to the DCCU to calculate the next DC simultaneously. Fig.4 shows the timing diagram when $\eta = 0, 2, 4$. We assume that the candidate polynomials initially have the constant terms only at the first iteration ($\eta = 0$). The value $q^v_{(1,r)}$ at the first iteration $\eta = 0$ denotes the updated coefficient in $Q_\eta(X,Y)$.

B. Discrepancy Coefficient Computation Unit (DCCU)

The DCC defined in equation (4) calculates the coefficient of the monomial $X^\alpha Y^\beta$ in $Q(X+x, Y+y)$ that is the shifted version of $Q(X,Y)$ by x and y in X, Y direction respectively, where α, β are non-negative integers that satisfy $\alpha + \beta < m$. The following equation shows the Hasse derivative [11] to find the DC.

$$d_v^{(\alpha,\beta)} = coef\left(Q_v(X+x, Y+y), X^\alpha Y^\beta\right) = \sum_{t=\beta}^{r} \sum_{s=\alpha}^{w_{v,t}} \binom{s}{\alpha}\binom{t}{\beta} q^v_{(s,t)} x^{s-\alpha} y^{t-\beta}, for\ v = 0, 1, 2, \dots, r \quad (7)$$

The hardware to solve equation (7) may suffer from latency, because it consists of a double loop of addition. The architecture proposed by Wang and Ma utilized the finite field additions, instead of multiplications, by representing the symbol by its exponent [6]. The proposed architecture calculates the DCs with respect to Y in parallel, whereas it calculates the DCs with respect to X and the candidate polynomials in serial. Equation (7) can be expanded as follows.

$$d_v^{(\alpha,\beta)} = \sum_{s=\alpha}^{w_{v,t}} \binom{s}{\alpha} x^{s-\alpha} \left\{ \binom{0}{\beta} q^v_{s,0} y^{0-\beta} + \binom{1}{\beta} q^v_{s,1} y^{1-\beta} + \cdots + \binom{r}{\beta} q^v_{s,r} y^{r-\beta} \right\} \quad (8)$$

The values $s-\alpha, t-\beta, (0 \le t \le r)$ are meaningful when $s > \alpha, t > \beta$. Fig.5 shows the block diagram of the proposed DCCU to compute equation (8). The multiplication at the input computes $x^{s-\alpha}$ and the DCC is performed monomially and in increasing order of X. Once the multiple of x, $x^{s-\alpha}$ is calculated, it is stored for the next computation and distributed to compute the other DCs in candidate polynomial simultaneously. In Fig.5, $y^{0-\beta}y^{1-\beta}, \dots, y^{r-\beta}$ denotes the output of the Y generator, to be explained later, and $q^v_{s,t}$, which is the output of the PUU, is the coefficient of monomial X^sY^t in the v-th candidate polynomial $Q_v(X,Y)$. The value $\binom{s}{\alpha}\binom{t}{\beta}$ can be easily implemented by Lucas's theorem [11]. $\binom{s}{\alpha}$ is the

common term like $x^{s-\alpha}$ that will be distributed and $\binom{t}{\beta}$ can be further simplified according to t. The registers on the right in Figure 5 store the intermediate DC values for each candidate polynomial.

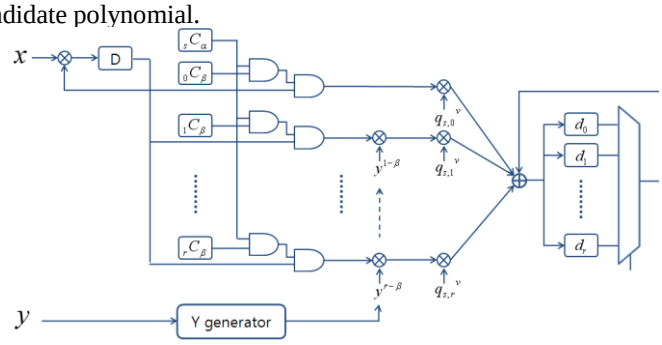


Figure 5. DCCU structure.

Y generator that is basically the same as that in [9] computes the multiple of y. When $\alpha = \beta = 0$, $y^{0-0} = y^0 = 1, y^{1-0} = y, \ldots, y^{r-0} = y^r$. At the first iteration, the value when $\beta = 0$ will be used immediately. As β increases, the values stored in the registers are shifted down and we can compute $y^{t-\beta}$ for the DCCU without any additional hardware. $r-1$ finite multipliers are required for the Y generator and the number of latency to get the first output is $\lfloor log_2 r \rfloor$.

C. Polynomial Update Unit (PUU)

The PU stage updates each candidate polynomial $Q_v(X,Y)$ using the selected "pivot" polynomial $Q_\eta(X,Y)$, where η is the index of the minimum weighted degree polynomial among all polynomials with non-zero DCs. As explained in Fig.2, the "non-pivot" polynomials are usually updated first and the "pivot" polynomial is updated last.

$$Q_v(X,Y) = \begin{cases} d_\eta^{(\alpha,\beta)} Q_v(X,Y) + d_v^{(\alpha,\beta)} Q_\eta(X,Y), & for\ v \neq \eta \\ Q_\eta(X,Y)(X+x), & for\ v = \eta \end{cases}$$

Here, $d_\eta^{(\alpha,\beta)}$ and $d_v^{(\alpha,\beta)}$ denotes the DCs of $Q_\eta(X,Y)$ and $Q_v(X,Y)$ respectively. The candidate polynomials for $v \neq \eta$, are updated by adding two polynomials: $Q_v(X,Y)$ multiplied by $d_\eta^{(\alpha,\beta)}$ and $Q_\eta(X,Y)$ multiplied by $d_v^{(\alpha,\beta)}$. This deletes the monomial $X^\alpha Y^\beta$ in $P(X,Y)$ in equation (3) that is the shifted version of the polynomial $Q(X,Y)$ by (x_i, y_i). Last, the polynomial $Q_\eta(X,Y)$ will be updated by multiplying $(X+x)$.

The proposed architecture processes the polynomials in serial but with an efficient schedule. The update for the "non-pivot" polynomials in case of $v \neq \eta$ requires the information of both $Q_v(X,Y)$ and $Q_\eta(X,Y)$ before update. The monomials with the same X degree in the polynomial are computed simultaneously and the update is preceded from the constant terms to higher order of monomials, since the proposed architecture processes the Y degrees in parallel.

We can serialize and rewrite the update procedure, as in equation (11).

$$q'_{i,y} = q_{i-1,y} + x q_{i,y}, 0 \le i \le d_x,\ 0 \le y \le r,\ q_{-1,y} = 0 \quad (11)$$

Figure 6 depicts the structure of one PUU element that can update both $Q_v(X,Y)$ and $Q_\eta(X,Y)$. When the corresponding coefficient $q_{i,y}^\eta$ in the polynomial $Q_\eta(X,Y)$ comes in as an input, the multiplexer select signal '*sel*' is set to '0' and the computation of $q_{i,y}^{\eta\,'} = x q_{i,y}^\eta + q_{i-1,y}^\eta$ is implemented. In the register, the coefficient $q_{i-1,y}^\eta$ will be stored to update the other polynomials $Q_v(X,Y)$. Then, this structure is used to update the polynomials $Q_v(X,Y)$ by setting the multiplexer select signal '*sel*' to '1', so the computation $q_{i,y}^{v\,'} = d_v q_{i,y}^\eta + d_\eta q_{i,y}^v$ will be implemented.

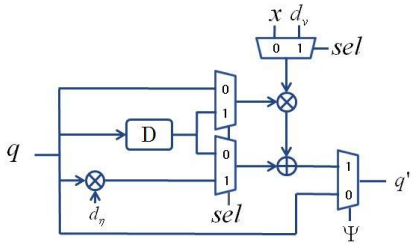


Figure 6. Structure of one PUU element.

Figure 7 shows the PUU structure that updates the polynomials in Y in parallel. PUU consists of $(r+1)$ PUU elements and each PUU element computes for the polynomial $\theta_v(X)$ in equation (2). The updated coefficients will be stored in the registers at the output and will be sent to the DCCU simultaneously to overlap the operations of the PUU and the DCCU. After finishing update in the registers, the data in the registers will be stored in the memory immediately, so the memory access occurs once every $(r+1)$ clock cycles. The number of memory accesses is reduced and the memory structure is simplified by applying the proposed scheduling.

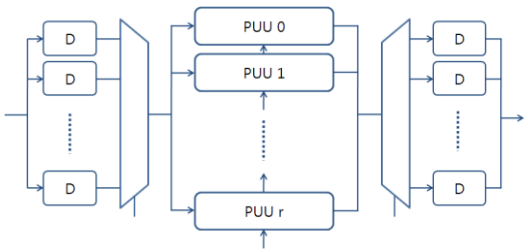


Figure 7. PUU overall structure.

D. Polynomial Order Sorting Unit (POSU)

In each iteration of the interpolation algorithm, the polynomial $Q_\eta(X,Y)$ needs to be selected by the weighted degree and the DC computed in the DCCU. Fig.8 shows the structure of the proposed POSU. Instead of computing the weighted degrees of the candidate polynomials to select $Q_\eta(X,Y)$ each iteration, we save the candidate polynomials with their weighted degrees once in the internal memory and

update the weighted degrees every iteration. The proposed POSU has $(r+1)$ registers that store the weight degree and its index in one word. As shown in equation (10), the degree of the polynomial $Q_\eta(X,Y)$ will be increased by one due to the multiplication by $(X+x)$, whereas the degrees of the other polynomials remain the same.

Figure 8 shows the POSU structure that reorders the polynomials by their weighted degrees. The registers consist of the $(r+1)$ shift registers and each register is partitioned to the part for the weighted degrees (O_v) and the part for the index of the polynomials (v), where O_v is initialized to $1,(k-1),2(k-1),\cdots,r(k-1)$ (k is message symbol, r is the number of the candidate polynomials) and v is initialized to $0,1,2,\ldots,r$. The weighted degree O_v is increased by one in case that $v=\eta$ and a bubble sort is performed to reorder the weighted degrees using comparator and shift registers.

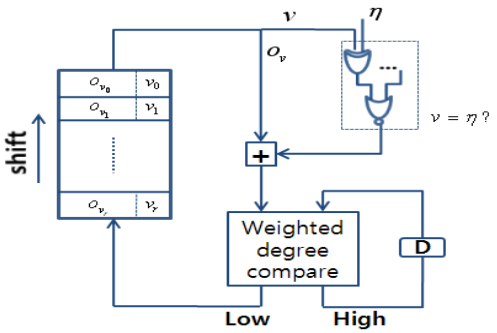


Figure 8. POSU structure.

IV. Design and Performance Analysis

In this section, we apply the proposed architecture to the $RS\ code\ (n,k)=(255,239)$ defined on $GF(2^8)$. We analyze the proposed architecture in terms of hardware cost, latency, and performance and compare it to existing architectures. The primitive polynomial used in this paper is $p(x)=1+x^2+x^3+x^4+x^5$. For fair comparison, we use the same experimental condition as that used in [6]. In the KV front-end, the maximum multiplicity, m_{max}, equals 5, using the low complexity method in the case of $\lambda=5$. The simulation shows the soft-decision decoder with an interpolation cost, C = 3,800, can provide more than 0.5dB of coding gain at a codeword error rate of 10^{-5} compared to the hard-decision decoder [11]. The following equation can be applied to estimate the number of bivariate polynomials [6].

$$r=min\left\{t\in Z:(t+1)\left(\frac{t(k-1)}{2}+k\right)>C\right\},$$

where Z denotes the set of all integers and C is the cost of the interpolation. By applying these parameters to this equation, r equals to 5. The re-encoding technique to reduce the number of iterations is used to achieve higher throughput. The number of iterations when the re-encoding is applied can be computed to $C'=C\times(n-k)/n=3{,}800\times16/255\cong239$. When C' is applied to the interpolation algorithm, the number of degrees of X in the polynomials can be computed to $C'/(r+1)=239/6\cong40$. The total number of latencies is $\left(\left\lceil\frac{C'}{2}\right\rceil+\varepsilon\right)C'$, where ε is the number of clocks to select and control the "pivot" polynomial, $Q_\eta(X,Y)$ in Figure 4.

TABLE I.
HARDWARE COMPLEXITY AND LATENCY OF DCCU AND PUU

Module		HW complexity	Latency
DCCU	$GF(2^q)$ multipliers	$3r+1$	$\left\lceil\frac{C}{2}\right\rceil$
	$GF(2^q)$ adders	$r+1$	
	registers	$(2r+3)+\lfloor log_2 r\rfloor$	
PUU	$GF(2^q)$ multipliers	$2(r+1)$	$\left\lceil\frac{C}{2}\right\rceil$
	$GF(2^q)$ adders	$(r+1)$	
	Registers	$(r+1)+2(r+1)^2$	
Total			$(\left\lceil\frac{C}{2}\right\rceil+\varepsilon)C$

Table I shows the hardware cost and latency of the DCCU and the PUU. The architecture in [5] uses the relatively complex dual-port memory. Also, the architectures in [5] and [6] divide the memory into $(r+1)^2$ of small memory modules that require a bigger area and more controls. We expect that the benefit of getting rid of complex access control is more than the degradation of power consumption and memory access speed. All the required coefficients can be read out and stored simultaneously and they are fed to the DCC and the PU in an efficient way. Also, by utilizing one large memory module with one-port, instead of multiple memory banks with dual-ports, the memory structure is simplified and efficient.

TABLE II.
HARDWARE COMPLEXITY AND PERFORMANCE COMPARISON

Design	Area (# of XOR gates	Critical Path (# of gates)	Latency	Throughput (normalized)	Efficiency (normalized)
[5]	8535	12	1437	1	1
[6]	11726	4	1775	2.43	1.77
[13]	7872	10	916	1.88	2.04
[14]	10718	12	454	3.17	2.52
Proposed	1321	4	10650	0.4	2.62

Table II compares the hardware complexity and the performance with the existing architectures. Ahmed, Koetter, and Shanbhag use a point-serial algorithm for the interpolation and apply a parallelism on polynomials and Y degrees [5]. The point-serial algorithm usually improves the performance when the interpolation points have high multiplicities. However, the hardware cost of the DCCU also increases due to τ $(\tau=2^{\lceil log_2 m\rceil})$, which is normally greater than r. The architecture proposed by Wang and Ma [6] also applies to a parallelism on polynomials and Y degrees and uses a hybrid data format for conversion between normal and power representations, so the computation complexity between symbols on finite field is dramatically reduced. However they need hardware for pre- and post-processing for the format conversion, such as a look-up table (LUT).

 ISBN: 978-1-61208-203-5

A finite field multiplier can be implemented by 64 XOR gates and 48 AND gates by employing composite field arithmetic, whereas a finite field adder simply requires 8 XOR gates. As analyzed in [6], the hardware complexity of the interpolation architecture grows linearly with $(m_{max}+1)^2$. For fair comparison, the proposed architecture including the architectures [5], [6] are scaled down with $m_{max}=2$ since m_{max} is equal to 2 in the BGMD architectures [13], [14]. All possible optimizations have been applied to the architectures in [5], [6]. Also, we can apply the pipelining to the proposed architecture for further speedup like the architecture in [6]. Based on that, the critical path can be reduced to the delay of 4 XOR gates. The hardware cost is analyzed based on the following assumption. Each AND gate or OR gate requires 3/4 of the area of an XOR, each MUX or memory cell has the same area as an XOR, and each register occupies about 3 times of the area of an XOR. According to Table I, the hardware complexity of the proposed architecture when $m_{max}=5$ is 5284 in equivalent XOR gates including memory. In case of $m_{max}=2$, the area requirement of the proposed architecture is equivalent to that of 1321 XOR gates. Since the terms of X degrees are processed in serial, the latency of the proposed architecture will be increased to the order of $(r+1)$, compared to the architecture in [6]. The analysis results for the existing architectures in Table II can be found in the paper [14]. Even though the throughput is relatively low, the efficiency is highest among the architectures in Table II.

TABLE III. RESULTS OF DESIGN AND SYNTHESIS

parameters				performance	
C	m_{max}	r	ε	total latency	Max clock freq.
239	5	5	4	29636	200 Mhz

	DCCU	PUU	control	total
gate count	7K	11K	7.1K	25.1K

The proposed interpolation architecture for the (255, 239) RS list decoder is designed with VerilogHDL and synthesized using a DongbuAnam 0.18 ㎛ standard cell library. Table III shows the synthesized gate count for the functional blocks in the proposed interpolation architecture. We use $C=239$, $m_{max}=5$, $r=5$, and $\varepsilon=4$ as the design parameters. The maximum operating clock frequency is 200MHz and the synthesized gate count is about 25.1K gates in two-input equivalent NAND gates.

V. CONCLUSIONS

In this paper, we proposed an efficient architecture with low hardware complexity for interpolation in soft-decision list decoding of Reed-Solomon codes. The proposed architecture has several advantages over the existing architectures in the following view points: 1) it employs parallel processing only for Y degrees in bivariate polynomial $Q(X,Y)$ and shares hardware modules, thus reducing the hardware complexity; 2) the schedule is adaptively adjusted according to the "pivot" polynomial computed at each iteration, so the irregular memory access problem is resolved; 3) the number of internal registers is reduced by processing the polynomial monomially; 4) scheduling minimizes the number of memory accesses and simplifies the memory structure by combining and storing data in memory, and the proposed architecture consists of one-port memory and one bank of memory and is efficient in area; 5) the DCCU and the PUU in the proposed architecture are overlapped in schedule, so the total latency is reduced. The proposed interpolation architecture for the (255, 239) RS list decoder is designed with VerilogHDL in a ModelSim environment. After logic synthesis, using the DonbuAnam 0.18um standard cell library, the maximum operating clock frequency is 200MHz and the synthesized gate count is about 25.1K gates in two-input equivalent NAND gates.

ACKNOWLEDGMENT

We are grateful to the IC Design Education Center that provided us with a design environment.

REFERENCES

[1] R. E. Blahut, Theory and practice of Error Control Codes, Addison-Wesley, Reading MA, 1983.

[2] V. Guruswami and M. Sudan, "Improved decoding of Reed-Solomon and algebraic-geometric codes," IEEE Trans. Inf. Theory, vol. 45, no. 6, pp. 1755-1764, Sep. 1999.

[3] R. Koetter and A. Vardy, "Algebraic soft-decision decoding of Reed-Solomon codes," IEEE Trans. Inf. Theory, vol. 49, no. 11, pp. 2809-2825, Nov. 2003.

[4] R. Koetter, J. Ma, A. Vardy and A. Ahmed, "Efficient interpolation and factorization in algebraic soft decision decoding of Reed-Solomon codes," in Proc. of IEEE Symp. On Info. Theory, 2003.

[5] A. Ahmed, R. Koetter, and N. Shanbhag, "VLSI architectures for soft-decision decoding of Reed-Solomon codes," in Proc. ICC, pp. 2584-2590, 2004.

[6] Z. Wang and J. Ma, "High-speed interpolation architecture for soft-decision decoding of Reed-Solomon codes," IEEE Trans. VLSI systems, vol. 14, no. 9, pp. 937-950, Sep. 2006.

[7] W. J. Gross, F. R. Kschischang, and P. Gulak, "Architecture and implementation of an interpolation processor for soft-decision Reed-solomon decoding," IEEE Trans. VLSI systems, vol. 15, no. 3, pp. 309-318, Mar. 2007.

[8] W. J. Gross, F. R. Kschsichang, R. Koetter, and P. G. Gulak, "A VLSI architecture for interpolation in soft decision list decoding of Reed-Solomon codes," in Proc. of IEEE Workshop on Signal Processing Systems, 2002.

[9] A. Ahmed, R. Koetter, and N. Shanbhag, "Systolic interpolation architectures for soft-decoding Reed-Solomon codes," in Proc. IEEE Workshop Signal Process. Syst., pp. 81-86, 2003.

[10] G. L. Feng and K. K. Tzeng, "A generalization of the Berlekamp-Massey algorithm for multisequence shift-register synthesis with applications to decoding cyclic codes," IEEE Trans. Inf. Theory, vol. 37, no. 5, pp. 1274-1287, Sep. 1991.

[11] H. Hasse, "Theorie der Hoheren differentiale in einem algebraischen funktionenkorper mit vollkommenem konstantenkorper bei beliebiger charakteristik," J. Reine. Ang. Math., vol. 175, pp. 50-54, 1936.

[12] V. C. da Rocha Jr., "Digital sequences and the Hasse derivative," in Communications Coding and Signal Processing, B. Honary, M. Darnell, and P. Farrell (Eds.), Communication Theory and Applications, John Wiley and Sons Inc., vol. 4, pp. 256-268, 1997.

[13] J. Zhu, X. Zhang, and Z. Wang "Backward interpolation architecture for algebraic soft-decision Reed–Solomon decoding," IEEE Trans. VLSI systems, vol. 17, no. 11, pp. 1602-1615, Nov. 2009.

[14] X. Zhang and J. Zhu, "High-throughput interpolation architecture for algebraic soft-decision Reed–Solomon decoding," IEEE Trans. Circuits and systems, vol. 57, no. 3, pp. 581-591, Mar. 2010.

 ISBN: 978-1-61208-203-5

Two-stage Wideband Class-E Power Amplifier in a 130 nm CMOS Process

Danish Kalim, Adel Fatemi and Renato Negra

Mixed-Signal CMOS Circuits, UMIC Research Centre
RWTH Aachen University, 52056 Aachen, Germany
Ph: +49-241-8027763, Fax: +49-241-8022199
kalim@umic.rwth-aachen.de, adel.fatemi@rwth-aachen.de, negra@ieee.org

***Abstract*—A wideband switching-mode power amplifier (SMPA) provides an optimum solution to cover multiple wireless standards with high efficiency and a high level of integration in a low-cost CMOS process. In this paper, a single-ended two-stage PA operating at 2.5 GHz in 130 nm CMOS technology is presented. The main-stage comprises a class-E PA based on finite DC-feed inductance, which provides high output resistance and can, therefore, cover a wide frequency range. A class-E driver with interstage matching is used to drive the main-stage PA in order to obtain large overall power gain with an optimum PA performance. The main and the driver stages are operated at 3.3 V and 1.3 V, respectively. Simulations indicate that the PA provides peak output power of 23.0 dBm and peak power added efficiency (PAE) of 64.0 %. From 2.15 GHz to 3.1 GHz, an output power of more than 20.5 dBm with a gain of 16.5 dB and PAE of more than 50.0 % are simulated.**

***Index Terms*—Wideband, class-E, load transformation network (LTN), switching-mode power amplifier (SMPA), power added efficiency (PAE).**

I. INTRODUCTION

The upcoming wireless technology is in motion to provide high data rate with wider coverage capabilities. From mobile equipment manufacturers perspective, this development brings several challenges since these advancements are linked to high power dissipation and high costs of the mobiles. Hence, it demands the design of high performance circuits with high efficiency to increase the battery life in a low cost CMOS process. The power amplifier (PA) is one of the most critical components of the RF front-end as it is the most power hungry element and defines the overall efficiency of a transceiver. Improving PA efficiency significantly impacts battery life and thus a comprehensive effort is being made to implement highly efficient CMOS PAs [1].

Switching-mode power amplifiers (SMPAs) are nonlinear class of PAs and have the capability to obtain high efficiency, ideally 100 % at RF and microwave frequencies. The active device is used as an ON/OFF switch such that for any time instant, a nonoverlapping voltage and current exist at the device output. Each class of SMPA, namely class-D [2], E [3], F [4] and F^{-1} [5] requires specific harmonic impedance termination at the output of the active device in order to prevent power dissipation. This eventually limits the operational bandwidth of a SMPA.

Introduction of next generation wireless communication standards increases the demand on both mobile equipments and base stations. A mobile terminal is required to cover the existing and also the upcoming standards to reduce the form factor and fabrication cost with minimum hardware effort. Several approaches for multiple band coverage have been proposed in literature. One solution is a multiband multiharmonic LTN using transmission-lines and/or multiple tuned LC circuits [6]- [7]. But, these are not feasible for mobile terminals due to their huge chip area requirements and the high associated cost.

The broadband design seems to be the most suitable approach to cover multiple bands which employs one active device as a switch and a wideband LTN [8]. It does not require any tuning element which results in an area efficient LTN and the PA can be fully integrated on-chip with a reasonable amount of die area.

A class-E PA based on finite DC-feed inductance instead of a conventional RF choke (RFC) provides wider bandwidth because it offers a higher optimum resistance, R_E, for the same output power and supply voltage [9]. In this paper, a two stage class-E PA using finite DC-feed inductance in 130 nm CMOS technology is presented. The main-stage is driven with a class-E PA to enhance the overall gain and PA efficiency. In simulation, the circuit provides 23.0 dBm peak output power with 64.0 % peak PAE. For a frequency range from 2.15 GHz to 3.1 GHz, the amplifier delivers an output power of 20.5 dBm and PAE of more than 50.0 %.

II. CLASS-E POWER AMPLIFIER

The class-E PA exploits the soft-switching property to provide high efficiency at high frequencies. It incorporates the drain-source capacitance, C_{DS} of the transistor in the LTN to eliminate the power losses associated with its discharging over the transistor in every RF cycle. This results in zero-voltage switching (ZVS) and zero-derivative switching (ZDS) which means at device turn on, there is no capacitor charge and no capacitor current at the device output [3].

 ISBN: 978-1-61208-203-5

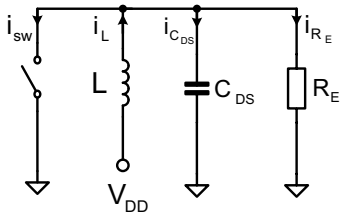


Fig. 1. Equivalent circuit of a class-E PA based on finite DC-feed inductance.

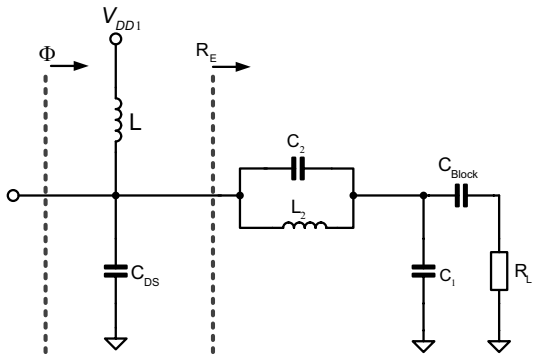


Fig. 2. Class-E LTN based on a finite DC-feed inductance and using the optimised second harmonic impedance termination network.

A. Load transformation network based on finite DC-feed inductance

The equivalent circuit of a class-E LTN using finite DC-feed inductance is shown in Fig. 1. The ZVS and ZDS conditions can be fulfilled by using only the LC_{DS} tank, whose resonant frequency is 1.41 times the operating frequency, f_0. The operation principle along with the equations for R_E, L and C_{DS} for a given output power, P_{out}, are detailed in [9]. The load phase angle, Φ, at the device drain at f_0 is given by:

$$\Phi = tan^{-1}\left(\frac{R_E}{\omega_0 L} - \omega_0 R_E C_{DS}\right) = 34.24^o\,, \qquad (1)$$

where impedances at all other harmonic frequencies are assumed to be capacitive. The LC_{DS} tank gives the accurate load angle and, thus, one can obtain the nominal class-E output waveforms at the transistor output. Since there is no imaginary part in the LTN as compared to the typical class-E conditions with an RFC, the optimum impedance can be written as:

$$Z_E(nf_0) = \begin{cases} R_E\,(1+j0)\,, & \text{if } n = 1 \\ \infty, & \text{if } n = 2, 3, .. \end{cases} \qquad (2)$$

A class-E PA using finite DC-feed inductance results in a relatively high R_E, which relaxes the impedance transformation ratio at f_0 and, therefore, also allows wider bandwidth. Secondly, the finite DC-feed inductor has a high self-resonance frequency (f_{SR}), which permits the designer to implement PAs for high frequency applications. In addition, the inductors can be integrated on-chip and, therefore, the total cost of the PA may be reduced.

B. Optimised second harmonic load circuit

The influence of harmonic termination on amplifier efficiency decrease with increasing order, whereas, second harmonic theoretically has the highest impact on efficiency. In order to filter out the second harmonic content, the parallel tank, L_2C_2, resonating at $2f_0$ is added to the circuit in Fig. 1. This leads to an optimised second harmonic load circuit for a class-E PA [10], as shown in Fig. 2.

The parallel tank L_2C_2 is inductive at f_0 and can be sized using LC_{DS} in combination with C_1 to get the nominal class-E conditions at f_0. As L_2C_2 resonates at $2f_0$, the value of L_2 is smaller than the series inductance used in a conventional class-E LTN. This is not only good for integration but it also extends the frequency range of the LTN. The values for the optimised second harmonic load circuit, illustrated in Fig. 2, are given by the expressions [10]:

$$C_1 = \frac{1}{2\pi f_0 R_L}\sqrt{\frac{R_L}{R_E} - 1}\,, \qquad (3)$$

$$L_2 = \frac{3R_E}{8\pi f_0}\sqrt{\frac{R_L}{R_E} - 1}\,, \qquad (4)$$

$$C_2 = \frac{1}{4(2\pi f_0)^2 L_2}\,, \qquad (5)$$

where R_L is the $50\,\Omega$ load.

III. Design and Implementation

Using the concept discussed in Section II, a two-stage class-E PA was implemented. The design starts with a class-E PA as a main-stage amplifier followed by a class-E driver and an interstage matching network.

A. Main-stage class-E PA

A class-E LTN with finite DC-feed inductance was designed using Fig. 2 for an output power, P_{out} = 0.7 W at 2.5 GHz. The transistor was biased at V_{DD_1} = 3.3 V and V_{GS_1} = 400 mV. This leads to an R_E of $21.5\,\Omega$, L of 0.93 nH and C_{DS} of 2.17 pF. The transistor was scaled to 2.3 mm for this design. The output capacitance of this transistor is smaller than the desired value, therefore, an external capacitance is added to provide nominal class-E conditions.

The fundamental impedance at the transistor drain is simulated to be $(27.0 + \mathrm{j}\,14.5)\,\Omega$, which corresponds to a load angle, $\Phi = tan^{-1}(\frac{14.5}{27.0}) \approx 28.2^o$, while other harmonics are capacitive. The achieved impedances at the fundamental and harmonic frequencies are in good comparison with the class-E switching conditions as depicted in Fig. 3.

Peak PAE of 64.8 % and peak output power of 23.6 dBm are simulated as shown in Fig. 4. The second and third harmonic frequencies are suppressed by 26.8 dBc and 36.6 dBc, respectively. Fig. 5 illustrates the simulated time-domain voltage and current waveforms at the device output, which approximate ideal class-E characteristics. In Fig. 6, amplifier performance

 ISBN: 978-1-61208-203-5

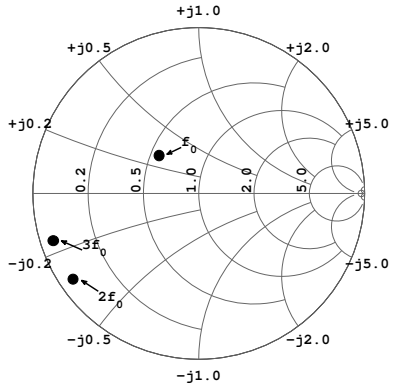


Fig. 3. Simulated output impedance of the designed class-E LTN based on finite DC-feed inductance.

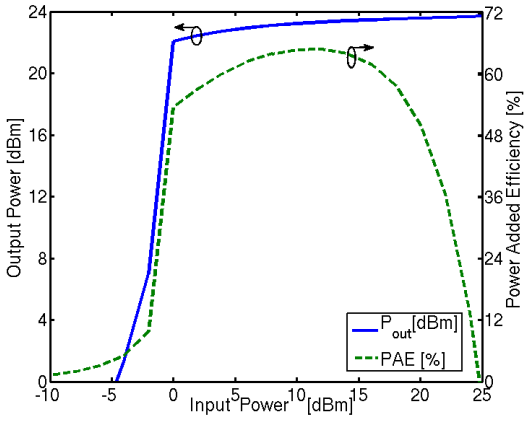


Fig. 4. Simulated output power and PAE of the designed class-E PA with finite DC-feed inductance.

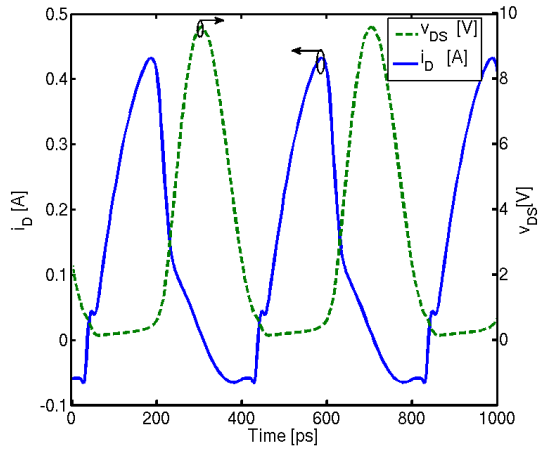


Fig. 5. Simulated time-domain current and voltage waveforms at the transistor's output of the class-E PA with finite DC-feed inductance.

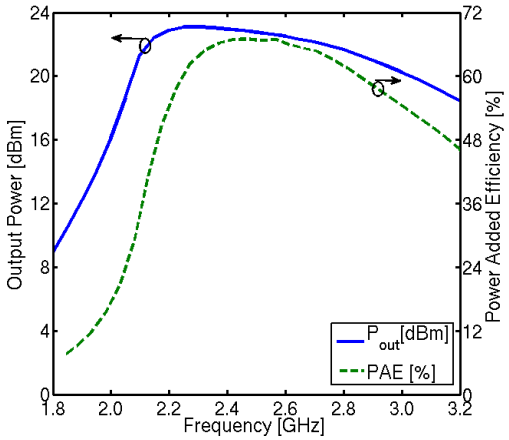


Fig. 6. Simulated output power and PAE versus frequency of the designed class-E PA with finite DC-feed inductance at P_{in} = 11 dBm.

against input frequency is highlighted. The main-stage PA attains more than 50.0 % PAE, more than 19 dBm of output power with an associated power gain greater than 8.0 dB from 2.15 GHz to 3.15 GHz.

B. Class-E driver stage

The driver stage consists of a class-E amplifier with an interstage matching network cascaded to the previously designed main-stage amplifier, as shown in Fig. 7. The class-E driver fulfills the class-E conditions as the tank L_3C_3 is tuned to $1.41f_0$. Secondly, the gate-source capacitance, C_{GS}, of the main-stage transistor is resonated out with L_4 at $2f_0$. It results in an approximately half-sinusoidal voltage waveform to operate the main transistor as a switch. Due to this, the PA operates only in the active region and the large negative swing of the input voltage is eliminated. Fig. 8 shows the simulated time-domain voltage waveforms at the output and input of the main transistor and also at the output of the class-E driver.

Simulated PAE and output power of the two-stage class-E PA are shown in Fig. 9. Peak PAE of 64 % and peak output power of approximately 23.0 dBm is obtained. One can clearly see that due to the driver and inter stage matching network, the amplifier performance is basically unchanged over a wide range of input powers. Fig. 10 illustrates the PA performance against input frequency. The designed amplifier provides PAE greater than 50 %, output power of more than 20.5 dBm and a power gain of more than 16.5 dB over a relative bandwidth of more than 36.2 %, i.e. from 2.15 GHz - 3.1 GHz.

Simulated performance of the SMPA is summarised in Table I. Following figure-of-merits (FOMs) are used to compare the performance of designed amplifier with other published two-stage CMOS SMPA designs:

$$FOM_1 = PAE \cdot Freq\ [\text{Hz}]^{0.25} P_{out}\,, \tag{6}$$

$$FOM_2 = PAE \cdot Freq\ [\text{Hz}]^{0.25} BW\,, \tag{7}$$

where BW is the relative bandwidth. In [12], inductors have large values which consume significant chip area. For this work, all the biasing inductors can be realised by bond wires, whereas an on-chip inductor, L_2 = 1.18 nH. Hence, the implemented two-stage amplifier has an area-efficient design.

 ISBN: 978-1-61208-203-5

TABLE I
PERFORMANCE COMPARISON OF THE PRESENTED DESIGN WITH OTHER TWO-STAGE CMOS SMPAS

Ref.	Tech. [nm]	V_{DD1} [V]	V_{DD2} [V]	Freq. [GHz]	Max Output [dBm]	Max PAE [%]	-3 % PAE, BW [GHz]	FOM_1	FOM_2
[11]*	180	2.0	-	1.9	16.3	70.0	1.87 - 1.96 (4.7 %)	6.28	6.87
[12]	180	2.5	1.8	2.4	23.0	73.0	2.22 - 2.57 (14.6 %)	32.3	23.6
This work	130	3.3	1.3	2.5	23.0	64.0	2.34 - 2.84 (19.3 %)	28.6	27.6

* Measurement

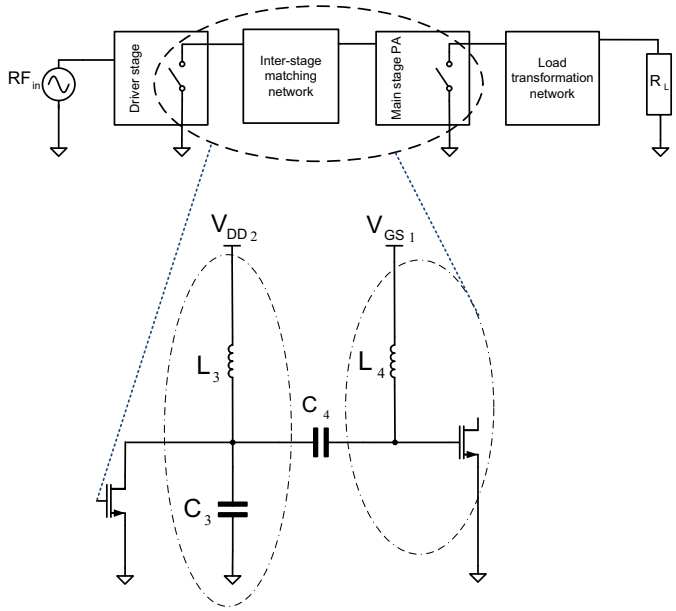


Fig. 7. Driver and interstage matching network for a two-stage class-E PA based on finite DC-feed inductance.

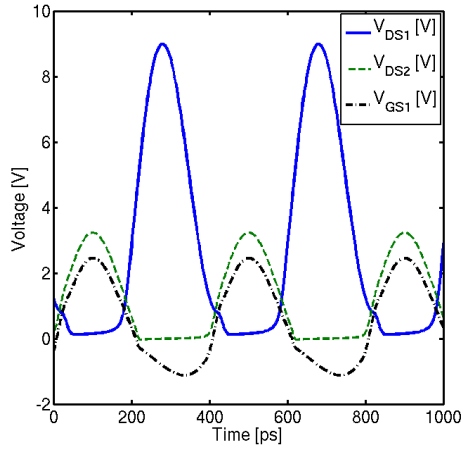


Fig. 8. Simulated time-domain voltage waveforms at the output and input of main-stage transistor and at the output of class-E driver.

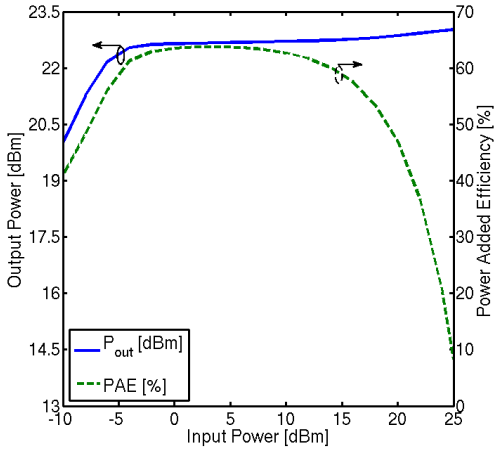


Fig. 9. Simulated output power and PAE of the designed two-stage class-E PA with finite DC-feed inductance.

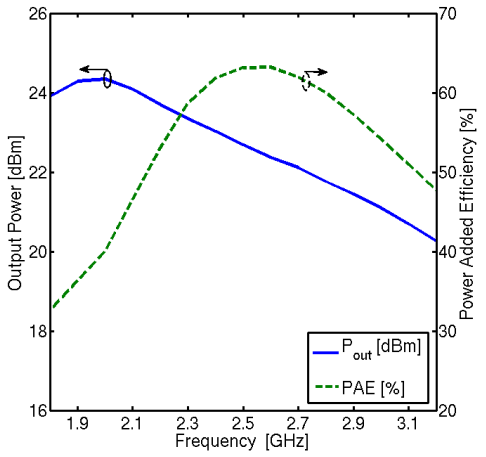


Fig. 10. Simulated output power and PAE versus frequency of the designed two-stage class-E PA with finite DC-feed inductance at P_{in} = 4 dBm.

IV. CONCLUSION

An efficient and wideband two-stage class-E PA in CMOS 130 nm process has been designed. The main-stage amplifier is based on class-E load network using finite DC-feed inductance. This approach leads to high optimum resistance, which facilitates wideband operation. Secondly, the use of finite DC-feed inductance has less resistive losses, high f_{SR} and, therefore, high efficiency. A class-E driver also using finite DC-feed inductance with interstage matching feeds the

 ISBN: 978-1-61208-203-5

main-stage class-E PA with a waveform approaching a half sinusoidal voltage. The benefit of the circuit is an improvement in the overall PA performance since there is no negative voltage swing.

Simulation results show that the amplifier can deliver 23.0 dBm peak output power at peak PAE of 64.0 % to a 50 Ω load at 2.5 GHz from 3.3 V supply. The two-stage PA obtains an optimum wideband performance with output power and PAE more than 20.5 dBm and 50.0 %, respectively, for a power gain greater than 16.5 dB, within a frequency range from 2.15 GHz to 3.1 GHz. The mentioned bandwidth covers WLAN, Bluetooth and LTE applications.

ACKNOWLEDGMENT

This work has been supported by the UMIC Research Centre, RWTH Aachen University.

REFERENCES

[1] Y. Ding and R. Harjani "A high-efficiency CMOS +22-dBm linear power amplifier," *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1895-1900, Sep. 2005.

[2] T. Nakatani, J. Rode, D. F. Kimball, L. E. Larson and P. M. Asbeck, "Digital polar transmitter using a watt-class current-mode class-D CMOS power amplifier," *IEEE J. Radio Frequency Integrated Circuits Symp.*, pp. 1-4, 2011.

[3] N. O. Sokal and A. D. Sokal, "Class E - A new class of high efficiency tuned single-ended switching power amplifiers," *IEEE J. Solid-State Circuits*, vol. 10, pp. 168-176, Jun. 1975.

[4] F. H. Raab, "Maximum efficiency and output of class-F Power amplifier." *IEEE Trans. Microw. Theory Tech.*, vol. 49, no. 6, pp. 1162-1166, Jun. 2001.

[5] Y. Y. Woo, Y. Yang and B. Kim, "Analysis and experiments for high-efficiency class-F and inverse class-F power amplifiers," *IEEE Trans. Microw. Theory Tech.*, vol. 54, no. 5, pp. 1969-1974, May 2006.

[6] D. Kalim and R. Negra, "Concurrent planar multiharmonic dual-band load coupling network for switching-mode power amplifiers," *Proc. IEEE MTT-S Int. Microw. Symp.*, pp. 1-4, Jun. 2011.

[7] R. Negra, A. Sadeve, S. Bensmida, F. M. Ghannouchi, "Concurrent Dual-band Class-F Load Coupling Network for Applications at 1.7 and 2.14 GHz," *IEEE Trans. on Circuits and Systems-II*, vol. 55, no. 3, pp. 259-263, Mar. 2008.

[8] A. Mazzanti, L. Larcher, R. Brama and F. Svelto, "A 1.4 GHz-2 GHz wideband CMOS class-E power amplifier delivering 23 dBm peak with 67 % PAE," *IEEE J. Radio Frequency Integrated Circuits Symp.*, pp. 425-428, 2005.

[9] A. V. Grebennikov and H. Jaeger, "Class-E with parallel circuit-A new challenge for high-efficiency RF and microwave power amplifiers," *IEEE MTT-S Dig.*, pp. 1627-1630, 2002.

[10] R. Negra and W. Bächtold, "Lumped-element load-network design for class-E power amplifiers," *IEEE Trans. Microw. Theory Tech.*, vol. 54, no. 6, pp. 2684-2690, Jun. 2006.

[11] C. -C. Ho, C. -W. Kuo, C. -C. Hsiao and Y. -J. Chan, "A fully integrated class-E CMOS power amplifier with class-F driver stage ," *IEEE Radio Frequency Integrated Circuits (RFIC) Symp.*, pp. 211-214, Jun. 2003.

[12] T. Wang, "Optimised class-E RF power amplifier design in bulk CMOS," *MSc.Thesis*, University of Texas, Arlington, Dec. 2007.

 ISBN: 978-1-61208-203-5

A SMART RFID Transponder

Riad Kanan
University of Applied Sciences
Rte du Rawyl 47
1950 Sion, Switzerland
riad.kanan@hevs.ch

Darko Petrovic
University of Applied Sciences
Rte du Rawyl 47
1950 Sion, Switzerland
darko.petrovic@hevs.ch

Abstract—**This paper presents a semi-passive universal, multi-applications, SMART RFID Sensing Transponder (SRST). It consists of a new low-power passive 13.56MHz RFID Analogue Front End (AFE), a micro-controller, sensors and rechargeable battery. The AFE was designed and fabricated successfully based on using a 0.35um CMOS technology. To allow re-charging a battery through the RF field, a new RF energy harvesting system is designed and integrated within the AFE; this leads to a self-powered system, which is a new benefit in the RFID sensing and continuous monitoring.**

Keywords-SMART RFID; sensing system; Low-Power.

I. INTRODUCTION

RFID (Radio Frequency Identification) technology has been widely used in the past few years as it helps identify objects and people in a fast, accurate and inexpensive way. It is used into many areas, including product tracing, transportation payment, animal identification, as well as passports, etc. [1].

RFID systems are comprised of three main components: the tag or transponder, the reader or transceiver that reads and writes data to a transponder, and in some applications, the computer containing database and information management software.

RFID tags can be active, passive, or semi-passive. The communication of active RFID is powered by its own battery which enables higher signal strength and extended communication range of up to 100 m. But the implementation of active communication requires larger batteries and more electronic components leading to higher costs. Passive and semi-passive RFID send their data by reflection or modulation of the electromagnetic field that was emitted by the Reader. The typical reading range is between 10cm and 3m.

In recent years, RFID has been introduced in sensing applications and semi-passive RFID tags are mainly used for such applications. The battery of semi-passive RFID is only used to power the sensor and recording logic. New developments have provided solutions for temperature monitoring, but RFID for sensing applications is still limited to sensing and storing the temperature and fulfilling the functionality of data logger [2][3]. Semi-passive RFID loggers offer an economical solution for the spatial profiling of transports with a high number of loggers.

Data loggers are standard tools for the supervision of cool chains. In order to handle the data for a high number of temperature records, the measurements have to be processed locally. It is not feasible to transmit full temperature data by a reader at unloading of a truck or container [4][5].

The shelf life prediction system is an example of application where huge data have to be processed locally. The system used to monitor the changes in quality of perishable foods during the transport phase and record them [5][6]. When addressed, the system delivers the prediction of the remaining shelf life time based on the used keeping quality model which uses the Arrhenius' Law, saying that, all reaction rate constants are assumed to depend on temperature [7][8].

An intelligent RFID sensing transponder has to measure the ambient parameters, process the information locally and transmits only the important information. Currently, there is no RFID transponder with a freely programmable processor. Indeed, the available RFID transponders do not allow the development of new applications because their protocols are already frozen by the chip logic core.

In this paper, a SMART RFID Sensing Transponder system is presented. It includes a new passive RFID chip and a platform which allows the following:

- A freely programmable processor for the development of new applications
- Cost-effective to facilitate the deployment
- The system allows storing the history of the measured environmental condition data (Temperature, Humidity, etc.)
- Layered HW/Firmware/SW system structure which allows a modular structure. The transponder provides an easy update and offers the possibility to extend application domains

Energy consumption of the transponder system has been minimized to a high extent by appropriate design and system control.

 ISBN: 978-1-61208-203-5

In addition, a Smart energy harvesting has been developed to ensure energy autonomy and to allow sensing and continuous monitoring. These represent radical progress in RFID sensing applications.

The paper is organized as follows. In Section II, we describe the proposed transponder system. In Section III, we present the transponder power consumption analysis. Section IV presents the designed Analogue Front End chip. Section V discusses the AFE experimental results. Finally, we present a conclusion of the work in Section IV.

II. RFID SENSING TRANSPONDER SYSTEM

The transponder system (SRST) is a semi-passive element. Figure 1 shows the general architecture of the tag which is designed with the minimum electronics components and optimized to achieve a low current consumption during operating period. The tag is composed of a new passive RFID Analogue Front End (AFE) chip, a micro-controller, EEPROM and sensors.

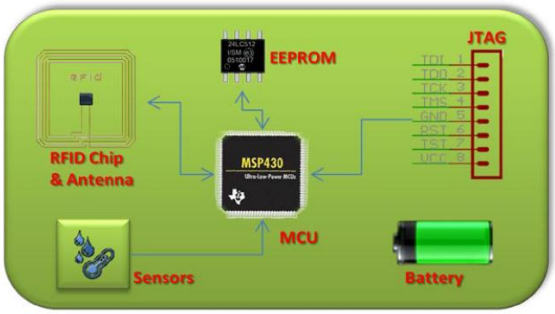


Figure 1 – SMART RFID Sensing Transponder architecture

Figure 2 and Figure 3 show the front and the back view of the realised prototype device for the transponder.

Figure 2 – SMART RFID Sensing Transponder prototype (Front view)

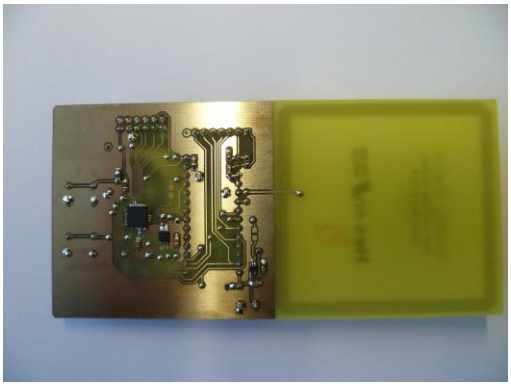


Figure 3 – SMART RFID Sensing Transponder prototype (Back view)

The component list used in the proposed sensing transponder is summarized in Table I. A microcontroller is used to develop applications and process data provided by sensors and stored in an EEPROM. Thanks to the new AFE intended mainly to the communication and energy scavenging, the SRST allows freely programmable processor for the development of new applications.

TABLE I. SLTT COMPONENT LIST

Component	Model	Notes
Microcontroller	MSP430F2350IRHA	16KB Flash, 256KB RAM
Sensor	SHT11	Temp & Humidity
EEPROM	24LC256-I/SM	256KB
Regulator	TPS78233	
AFE	***Proprietary***	***13.56MHz RFID Analogue Front End***

To add sensors to the tag, a microcontroller becomes necessary for the analysis of the measured values. The processor module calculates and stores the resulting values into a programmable memory EEPROM. The stored data are sent when receiving a "data-log" command from the reader. The RFID AFE Chip transmits the stored data over the RF Field to the reader. In addition, the RFID tag sends a real time sensor measures. In this case, the microcontroller sends the measured values of the sensors, after a conversion, directly to the RFID chip. For the programming of the micro-controller and the development of a new application, a JTAG (Joint Test Action Group) interface is added.

III. TRANSPONDER OPTIMIZED POWER CONSUMATION CONSIDERATION

The transponder system is optimized for low-power dissipation in order to extend the battery life, which allows the condition monitoring during the transport for instance. After the good shipment, a new battery charging cycle is started to prepare the transponders for a new shipment.

 ISBN: 978-1-61208-203-5

Table II summarizes the power consumption of each component of the transponder.

TABLE II. POWER CONSUMPTION SUMMARY

Chip	Active mode	Standby mode
MSP430F2330	390µA @ 3V, 1MHz	1µA @3V
SHT11	550µA @3.3V	0.3µA @3.3V
24LC256	Read=400uA, Write=2mA @3V	5µA @5.5V

Extending the battery lifetime requires an efficient use of the transponder components which must then be turned off when not required. This is what is called the duty-cycling, i.e. the sensor is active during a short period and goes to the standby mode when its sensing information are sent and or stored in the EEPROM. The EEPROM is switched to the standby mode after a write cycle.

Figure 4 presents the current consumption at different operating steps of the transponder during one sampling period.

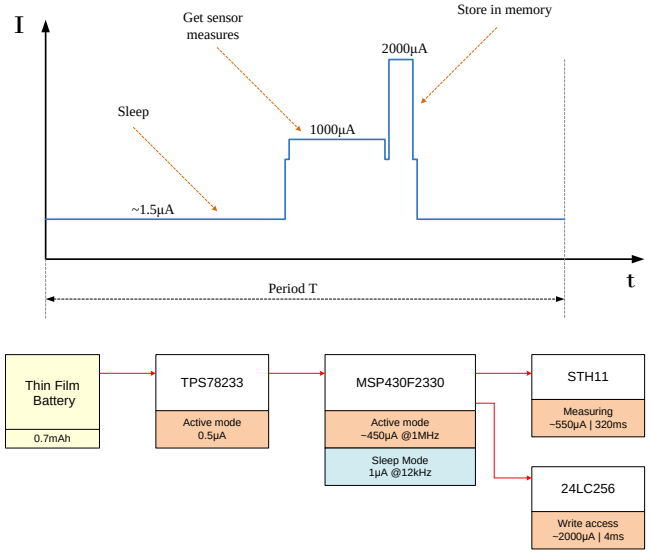


Figure 4 – Current consumption at different transponder operating steps

From Figure 4, the average current/Period (T) can be given by (eq. 1):

$$I_{avg} = \frac{1000uA \bullet 320ms + 2000uA \bullet 4ms + (T - 324ms) \bullet 1.5uA}{T} \quad (1)$$

For a battery with 0.7mAh capacity, the battery lifetime is then given by (eq. 2):

$$W_{time} = \frac{700uA \bullet 3600s}{I_{avg}} \quad (2)$$

Figure 5 presents the battery lifetime for different battery capacities as function of the sampling period. We can see for example, if a measurement period is set to 10mn, a battery with a capacity of 1mAh can been used for about 20 days before starting a new battery charging cycle. A small battery capacity is selected to get a thin form factor for the overall transponder.

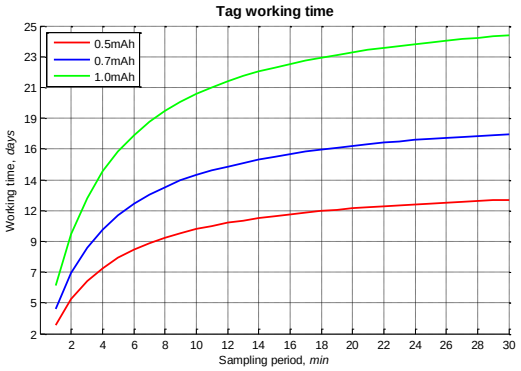


Figure 5 – Battery lifetime as function of the sampling period

IV. ANALOGUE FRONT END (AFE)

To allow free programing RFID transponder, a new RFID Analog Front End (AFE) has been designed and fabricated. Figure 6 is showing the schematic of the AFE.

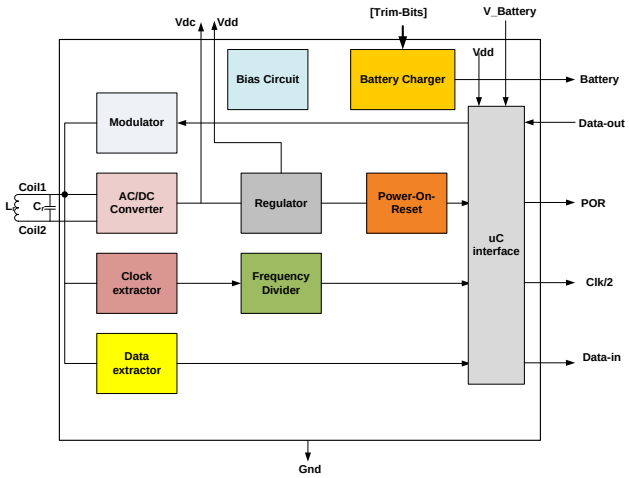


Figure 6 – RFID Analogue Front End Chip (AFE)

The AFE includes the following blocks:

A. *Power Supply*

The on chip power supply is extracted from the exciting field using an AC/DC converter. To avoid over-voltages in high magnetic fields the DC-voltage is clamped. The buffered Supply Voltage passes via a regulator. The output of the regulator is used to power the major part of the analogue front end.

The conventional rectifier is a diode bridge rectifier. The structure of bridge rectifier and its MOS construction are shown in Figure 7.

 ISBN: 978-1-61208-203-5

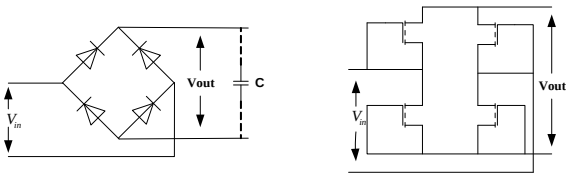


Figure 7 – Diode and MOS bridge rectifier

The diode bridge rectifier output voltage is given by Vout=Vin-2V_{th}.

In the MOS transistors bridge rectifier, the voltage drops are related to the threshold voltage and also the overload voltage, which linearly increases with square root of the current (eq. 3):

$$\Delta V = V_{TH} + \sqrt{\frac{2 \cdot L \cdot I}{C_{ox} \cdot W \cdot \mu}} \qquad (3)$$

To reduce the output voltage drop, the bridge rectifier structure shown in Figure 8 is used. In this structure, the output voltage drop is reduced from two threshold voltages to one threshold voltage.

In Figure 8, NMOS MN1 and MN2 are diode-connected structure, and NMOS MN3 and MN4 are cross-connected structure. L1 and L2 are the input differential signal, when L1 is high, L2 is low, MN1 and MN3 transistors are turn-on, MN2 and MN4 transistors are turn-off and the rectifier has one V_{GS} drop. The opposite operation occurs when L1 is low and L2 is high.

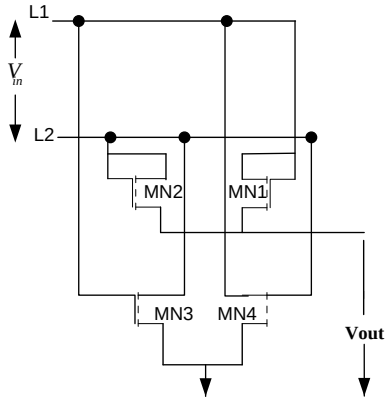


Figure 8 – Cross-connected MOS bridge rectifier

B. Power On Reset

The Power-On-Reset (POR) circuit monitors the regulated voltage V_{dd} and generates a global reset-signal putting the chip into an appropriate initial state at power up. It also will guarantee that the chip ceases operating when the supply voltage falls below level necessary for reliable operation. Hysteresis system is provided to avoid improper operation at the limit level

C. Modulator

The Modulator will modulate the continuous wave RF signal coming from the reader by changing the Q-factor of the tuned circuit by means of an extra resistive load connected in parallel with the resonance capacitor C_r. These changes in the Q factor induce a corresponding signal in the reader coil.

D. Clock Extractor

The clock extractor generates a system clock with the frequency of the RF field. The output signal of this Clock Extractor is passed via a Frequency divider and will then define the on-chip timings.

E. Data-Extractor

The data extractor demodulates the incoming signal to generate logic levels and decodes the incoming data.
In the ISO 15693 standard, communication between the reader and the tag takes place using the modulation principle of Amplitude Shift Keying (ASK). Two modulation indexes are used, 10% and 100%. The tag shall decode both. The reader determines which index is used. Data transmission type from tag to reader employs load modulation. When 100% ASK is selected, there is the discontinuousness in the energy of electromagnetism field because of characteristics of modulation type. When 10% ASK is selected, the transmission of energy of electromagnetism field is continuous.

Figure 9 and Figure 10 show the incoming signal and the extracted data in 10% and 100% modulation receptively.

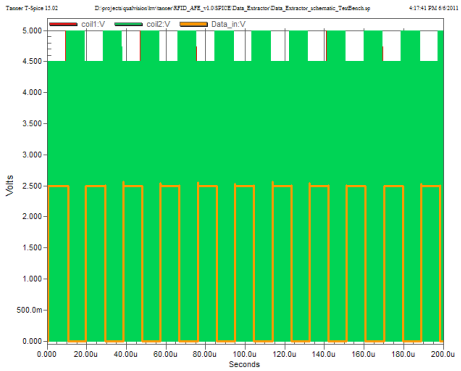


Figure 9 –10% OOK Modulation, Data extraction

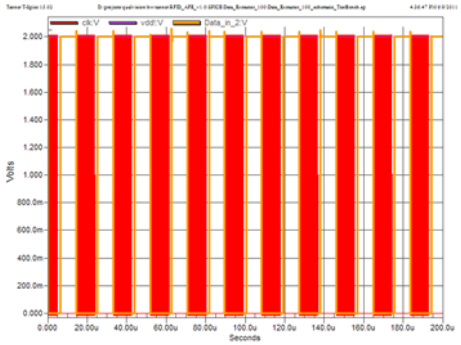


Figure 10 –100% OOK Modulation, Data extraction

 ISBN: 978-1-61208-203-5

F. *Micro-controller interface*

This block allows setting the appropriate voltage levels for the data coming in and out of the AFE.

G. *RF Energy harvesting*

The important feature of the proposed system is the re-use feature thanks to the battery charger block which will allow re-charging the battery through the RF field.

The battery charger has been designed to allow the charging of a Micro-Energy Cell (MEC®), which is a solid-state, rechargeable thin-film battery. MECs are manufactured by Infinite Power Solutions using wide area thin-film deposition techniques similar to those used to manufacture semiconductors [9]. The MECs enjoy the advantages of rapid recharge and charge acceptance at currents as low as 1uA.

Figure 11shows the block diagram of the battery charger. It consists of a battery current charger and a voltage sensing blocks.

The battery supply (V_{dc}) is obtained by rectifying the power carrier of the wireless link.

The battery current charger consists of reference current and current sources.

The voltage sensing block senses the battery voltage and generates the end-of-charge signal (Charge) so the microcontroller will set the Enable (uc_EN) low to stop the battery charging.

The selection of the current charge level and the battery end-of-charge is done by a trimming method. The trimming circuit was chosen by means of a binary weighted switch network with 11 bits resolution.

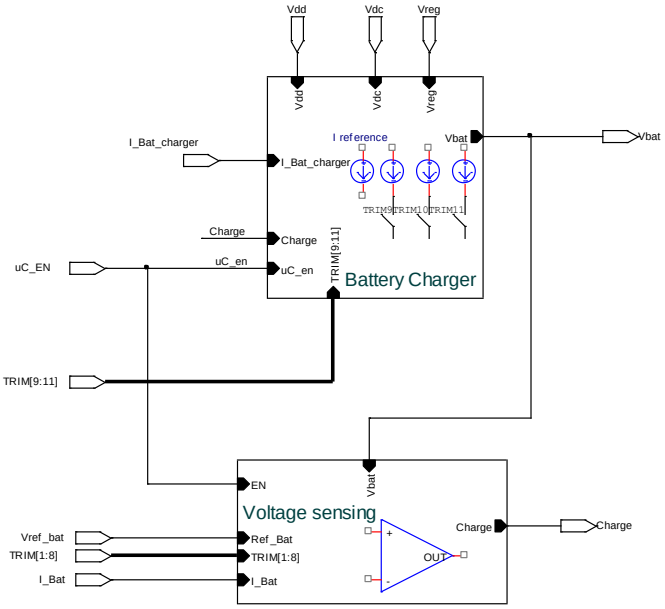


Figure 11 –Block diagram of the battery charger

V. AFE EXPERIMENTAL RESULTS

The proposed AFE has been designed and fabricated successfully using a 0.35um CMOS technology as shown in Figure 12. The overall circuit is 1635um by 1640um.

Before the layout release, every function module of the AFE has been verified by SPICE simulations and by considering the variation of Process, Voltage and Temperature (PVT).

The measured total current consumption of 6uA has been achieved in active mode. The AFE chip operates within a temperature range of -40 to 95°C and a supply range (V_{dc}) of 3-5 V.

A summary of the chip characterization is presented in this section.

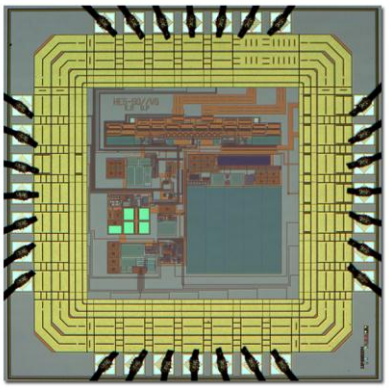


Figure 12 –Die photo of the AFE

A. *Reader-Transponder communication*

The SMART RFID Sensing Transponder communication platform is fully compliant with the standard protocol ISO15693 [10].

Custom commands for sensing and monitoring have been also developed and implemented.

Figure 13 is showing an example of communication. The reader sends Inventory request Double Sub-carrier, High Data-rate. The data are extracted for the RF field by the AFE (top trace). The transponder responds through the AFE (bottom trace) by sending its UID.

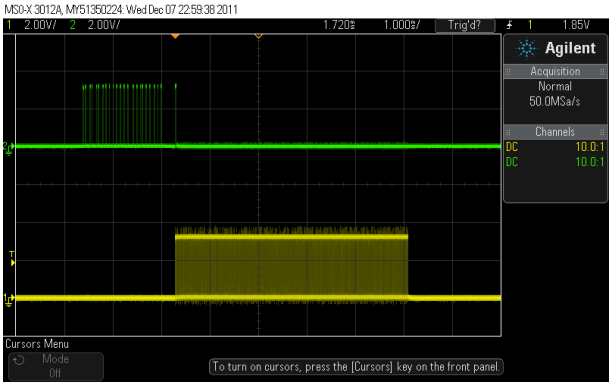


Figure 13 –Reader-Transponder communication: Inventory request

In Figure 14, the reader sends write command (top trace), the transponder responds successfully after 4.5ms (bottom trace).

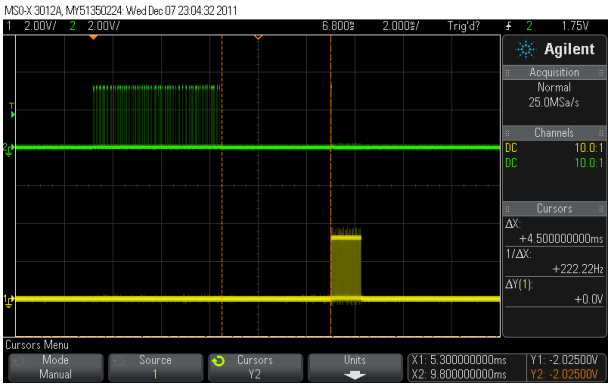


Figure 14 –Reader-Transponder communication: Write command

B. *RF enrergy harvesting*

To evaluate the battery charger, a capacitor of 100uF was used. The limit test cases are shown below:

a) Case 1

The battery voltage is set to the minimum level of 3.0V with the trimming bits 1 to 8 at low state. The charging current is minimum by setting the trimming bit 9 to 11 to low state (Figure 15).

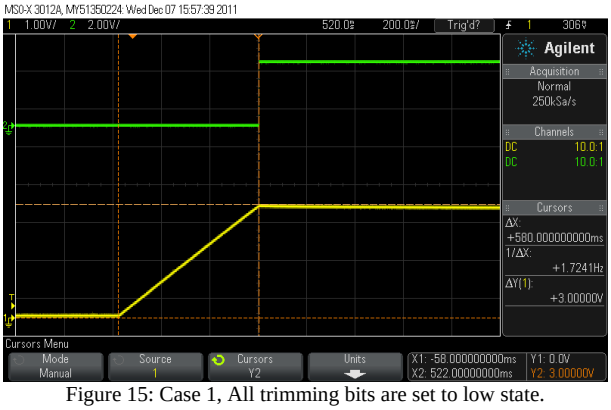


Figure 15: Case 1, All trimming bits are set to low state.
TRIM<1:8>= Low, TRIM<9:11>= Low Battery voltage = 3.0V,
Charge Time: 580ms

b) Case 2

The battery voltage is set to the maximum level of 4.1V with the trimming bits 1 to 8 at high state. The charging current is increased by setting the trimming bit 9 to 11 to high state (Figure 16).

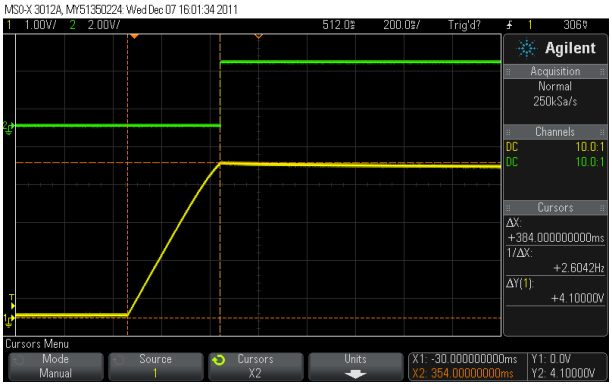


Figure 16: Case 1, All trimming bits are set to high state.
TRIM<1:8>= High, TRIM<9:11>= High, Battery voltage = 4.1V,
Charge Time: 384ms

VI. CONCLUSION

In this paper, a novel RFID sensing technology that is validated in a Smart, Self-powered, Low-cost and Re-usable transponder system was presented. A batteryless RFID Analogue Front End has been described. The chip was fabricated in a 0.35m CMOS process. The re-use and continuous features are allowed thanks to the designed RF energy harvesting system. In addition, low-power consumption has been achieved by optimizing circuit design and technology.

The power consumption of the used sensors is still high. In future, to further reduce the overall power consumption, low-power temperature and humidity sensors will be designed and integrated within the AFE.

ACKNOWLEDGMENT

This work has been supported by the HES-SO (Project No. 21958). The authors would like to thank the RCSO ISYS scientific committee.

REFERENCES

[1] K. Finkenzeller, “RFID Handbook: Fundamentals and Applications in Contactless Smart Cards and Identification,” 2nd Ed, Wiley, 2003.

[2] www.ksw-microtec.com [retrieved: May, 2012].

[3] www.turbo-tag.com [retrieved: May, 2012].

[4] R. Jedermann, W. Lang, “The minimum number of sensors-Interpolation of spatial temperature profiles,” Wireless Sensor Networks, 6th European Conference, Vol. 5432, pp. 232-246, February, 2009.

[5] R. Jedermann. J.P. Edmond. W. Lang, “Shelf life prediction by intelligent RFID,” Dynamics in Logistics. First International Conference, pp. 231-238, August, 2007.

[6] R. Jedermann, K. Stein, M. Becker, W. Lang, “UHF-RFID in the Food Chain – From Identification to Smart Labels,” 3rd International Workshop on Cold chain Management, pp. 3-15, June, 2008.

[7] L. Tijskens, J. Polderdijk, “A generic model for keeping quality of vegetable produce during storage and distribution,” Agricultural Systems 51(4), pp. 431-452, August, 1996.

[8] L. M. M. Tijskens, R.E. Schouten, “Modeling quality attributes and quality related product properties, ” Postharvest Handling: A Systems Approach, ISBN 978-0-12-374112-7, pp. 483-512, January, 2009.

[9] www.InfinitePowerSolutions.com [retrieved: May, 2012].

[10] Air interface and initialization, ISO/IEC 15693-2, 2000.

 ISBN: 978-1-61208-203-5

Modeling and Performance Evaluation of Small Cell Wireless Networks with Base Station Channels Breakdowns

Nawel Gharbi
Computer Science Department
University of Sciences and Technology, USTHB
Algiers, Algeria
Email: ngharbi@wissal.dz

***Abstract*—The ever-increasing number of customers and the need for higher data rates and multimedia services require the deployment of Small Cell Networks. This paper considers modeling, performance evaluation and reliability of Small Cell Wireless Networks, taking into account the retrial phenomenon, finite number of customers served in a cell and channels breakdowns. The aim of this paper is to give a detailed performance and reliability analysis of these next-generation networks considering different breakdowns disciplines using Generalized Stochastic Petri nets formalism. The novelty of this investigation consists in the consideration of two breakdowns disciplines: channels breakdowns and base station (synchronous) breakdowns. For the first one, each channel is an independent working unit, and it can fail independently of other channels state. For the second one, the breakdowns are synchronous, hence all the channels of the base station fail down simultaneously. Hence, we show how this high level model allows us to cope with the complexity of such finite-source retrial networks, under the different breakdowns disciplines and how several steady-state performance and reliability indices can be derived. Through numerical examples, we discuss the effect of the network parameters on performance.**

Keywords-Small Cell Networks; Retrial phenomenon; Channels breakdowns; Base station breakdowns; Generalized Stochastic Petri nets; Performance and reliability indices.

I. INTRODUCTION

The ever-increasing number of customers and the need for higher data rates and multimedia services lead to stringent requirements on the bit rate/km2 that next-generation cellular wireless networks are expected to deliver. A promising approach to solving this problem is through the deployment of Small Cell Networks (SCNs), which represent a novel networking paradigm based on the idea of deploying short-range, low-power, and low-cost base stations (BSs) operating in conjunction with the main macro-cellular network infrastructure. Small Cells operate in licensed and unlicensed spectrum that have a range of 10 meter to 200 meters, compared to a mobile Macrocell which might have a range of a few kilometers. The use of SCNs is envisioned to enable next-generation networks to provide high data rates, allow offloading traffic from the macro cell and provide dedicated capacity to homes, enterprises, or urban hotspots. SCNs encompass a broad variety of cell types, such as micro, pico, femto cells, as well as advanced wireless relays and distributed antennas. Regarding compatible technologies, Small Cells are available for a wide range of air interfaces including GSM, CDMA2000, TD-SCDMA, W-CDMA, LTE and WiMax.

This paper considers modeling, performance evaluation and reliability of small cell wireless networks, taking into account the repeated calls of customers and channels breakdowns. Models with repeated calls (or retrial phenomenon) arise in various practical areas as telecommunication, computer networks and cellular mobile networks [1], [2], [3]. These models are based on the fact that, when servers are all busy or unavailable, customers attempting to get a service are not put in a queue but will try again to reach the servers after a random delay. Significant references reveal the non-negligible impact of repeated calls on the network performances. These repeated calls arise due to a blocking in a network with limited capacity resources or are due to impatience of customers. For a recent summary of the fundamental methods, results and applications on this topic, the reader is referred to [4], [5], [6].

To this end, we observe a wireless network where a supported area is divided into small cells, each of them is served by a base station having a limited number of channels which could be subject to breakdowns. These random breakdowns may have a heavy influence on the network quality of service. On the other hand, the number of mobiles (or customers) served in a cell is also small, such that models with a finite number of sources should be considered. These three aspects, customer retrial, finite number of sources and breakdowns of the base station channels, will be dealt with in this paper.

Although the reliability study is of great importance, there are only few works that take into consideration retrial phenomenon involving the unreliability of the servers, as it can be seen in the recent classified bibliography of Artalejo [6]. Moreover, most studies deal with single unreliable server retrial queueing systems [7], [8] or an infinite customers source [9].

Regarding finite-source retrial systems with unreliable multiple servers, we have found some few papers as [10]

 ISBN: 978-1-61208-203-5

in which the servers are asymmetric (heterogenous) and the models are analyzed by queueing theory, and our recent paper [11] where retrial systems with servers breakdowns policy were analyzed using Generalized stochastic Petri nets (GSPNs) formalism. However, the several breakdowns mechanisms considered in the literature, can be classified as *servers breakdowns*.

In this paper, we propose the applicability of GSPNs for modeling and performance evaluation of Small Cell Networks (SCNs) with unreliable base station channels. The novelty of this investigation consists in the consideration of two breakdowns policies: servers (channels) breakdowns and station breakdowns. For the first one, each channel is an independent working unit, and it can fail independently of other channels state. For the second one, the breakdowns are synchronous, hence all the channels of the station fail down simultaneously (base station breakdown). This phenomenon occurs in practice, for example, when a system consists of several interconnected machines that are inseparable, or when all the machines are run by a single operator which may be fails at any time. In such situations, the whole station has to be treated as a single entity.

The paper is organized as follows: First, we give an overview of syntax and semantics of GSPNs formalism. In Section 3, we present the mathematical model describing the customers behavior in SCNs. Next, the GSPNs models for the different breakdowns disciplines are developed. In Section 5, we show how several steady-state performance and reliability indices can be derived. Then, based on numerical examples, we validate the proposed models with respect to the reliable case and we discuss the effect of the network parameters on performance. Finally, we give a conclusion.

II. Syntax and Semantics of Generalized Stochastic Petri Nets

In this section, we developed briefly the basics concepts of Generalized Stochastic Petri Nets formalism (GSPNs), that the readers are needed to better understand the proposed models describing small cell networks.

Generalized stochastic Petri nets [12], [13] are mathematical and graphical models, that are well suited for representing and analyzing stochastic and concurrent systems with synchronization characteristics.

A GSPN is a directed graph that consists of two kinds of nodes, called places (drawn as circles) and transitions that are partitioned into two different classes: timed transitions (represented by means of rectangles) describe the execution of time consuming activities and can fire only after a random delay characterized by a negative exponential probability distribution, and immediate transitions (represented by thin bars), which model logic activities as synchronization, have priority over timed transitions and fire in zero time once they are enabled. Formally, a GSPN can be defined as a seven-tuple $(P, T, I, O, Inh, M_0, W)$ where:

- P is the set of places;
- T is the set of timed and immediate transitions;
- $I, Inh : P \times T \rightarrow IN$ are the input and inhibitor functions, which provides the multiplicities of the input and inhibitor arcs from places to transitions (IN is the set of natural numbers);
- $O : T \times P \rightarrow IN$ is the output function which provides the multiplicities of the output arcs from transitions to places;
- $M_0 : P \rightarrow IN$ is the initial marking, which describes the initial state of the system;
- $W : T \rightarrow IR^+$ is a function that associates rates of negative exponential distribution to timed transitions and weights to immediate transitions.

An inhibitor arc is represented by a line terminating with a rounded head. The presence of a token in the inhibitor place inhibit the firing of the transition.

The system state is described by means of markings. A marking is a mapping from P to IN, which gives the number of tokens in each place. A transition is said to be enabled in a given marking, if and only if each of its normal input places contains at least as many tokens as the multiplicity of the connecting arc, and each of its inhibitor input places contains fewer tokens than the multiplicity of the corresponding inhibitor arc.

The firing of an enabled transition creates a new marking (state) of the net. The set of all markings reachable from initial marking M_0 is called the *reachability set*. The *reachability graph* is the associated graph obtained by representing each marking by a vertex and placing a directed edge from vertex M_i to vertex M_j, if marking M_j can be obtained by the firing of some transition enabled in marking M_i. This graph consists of *tangible markings* enabling only timed transitions and *vanishing markings* in which at least one immediate transition is enabled. Since the process spends zero time in the vanishing markings, they are eliminated from the reachability graph by merging them with their successor tangible markings [12]. This elimination results in a *tangible reachability graph*, which is isomorphic to a continuous time Markov chain (CTMC). The solution of this CTMC at steady-state is the stationary probability vector π which can be expressed as the solution of the linear system of equilibrium equations $\pi.Q = 0$ with the normalization condition $\sum_i \pi_i = 1$, where π_i denotes the steady-state probability that the process is in state M_i and Q is the infinitesimal generator. Having the probabilities vector π, we can compute several stationary performance indices of the system.

III. Mathematical Model

The mathematical model describing the customers behavior in small cell wireless networks can be viewed as a

retrial model consisting of a multiserver service station, an imaginary waiting space called *orbit* and a finite source of K homogeneous customers. Each customer can be in one of the following states: free, under service or in orbit at any time. The probability that any particular customer generates a primary request for service in any interval $(t, t + dt)$ is $\lambda dt + o(dt)$ as $dt \to 0$ if the customer is free at time t. The service station consists of c identical (symmetric) servers subject to breakdowns and repairs. Each server can be in operational (up) or non-operational (down) state, and it can be idle or busy. If one of the servers is *up and idle* at the moment of the arrival of a call, then the call starts to be served immediately, the customer moves into the under service state and the server moves into busy state. The service times are independent and identically exponentially distributed with parameter μ. After service completion, the server becomes idle. Otherwise, if all the servers are busy or down at the arrival of a call, the customer joins the orbit and starts generation of a flow of repeated calls exponentially distributed with rate ν until it finds one operational free server.

A server can fail during the interval $(t, t + dt)$ with probability $\delta dt + o(dt)$ as $dt \to 0$ if it is idle, and with probability $\gamma dt + o(dt)$ if it is busy. In the literature, three breakdowns disciplines were defined:

- The **active breakdowns discipline**[10], [9] when $\delta = 0$ and $\gamma > 0$. This means that a server can fail only in busy state.
- The **independent breakdowns discipline**[10] when $\delta > 0$, $\gamma > 0$ and $\delta = \gamma$. In this case, a server can fail in busy or free state with the same probability.
- The **dependent breakdowns discipline** which we have proposed recently in [11]. In this case, the failure probability depends on the server state. Hence, the rates $\delta > 0$ and $\gamma > 0$ could be equal or not.

If the server fails in busy state, the interrupted customer returns to the orbit to resume service later. The repair time of a server is exponentially distributed with a finite mean $1/\tau$. We assume that the repairman follows FIFO discipline to fix up the servers breakdowns, repairs one server at a time and after repair, the server is as good as new.

The several breakdowns mechanisms studied in the literature, can be classified as *servers breakdowns*. In this paper, we introduce the *station breakdowns policy*, which describes systems with synchronous breakdowns of all servers of the station.

IV. GSPN Models of Small Cell Networks with Channels and Base Station Breakdowns

In the following, we present the GSPN models describing Small Cell Wireless Networks with different breakdowns disciplines.

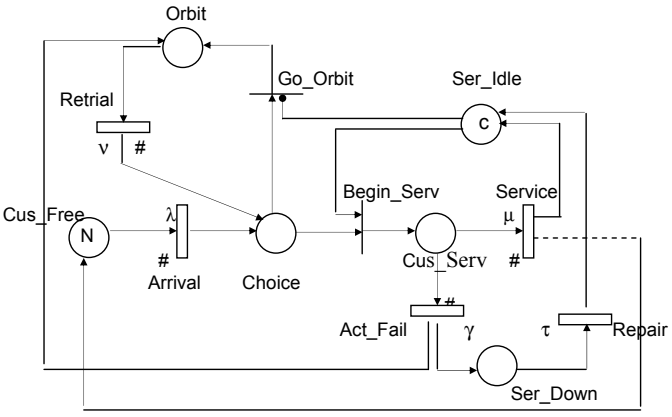


Figure 1. GSPN model of small cell wireless networks with retrials and active channels breakdowns

A. Retrial Networks with Active Channels Breakdowns

Figure 1 shows the GSPN describing a wireless network small cell with retrials and active breakdowns of channels. In this model, the place Cus_Free contains the free customers, the place $Orbit$ represents the orbit, Cus_Serv contains customers under service, Ser_Idle represents the operational free servers (channels) and the place Ser_Down contains non-operational (failed) servers.

The initial marking of the net is: $\{K, 0, 0, c, 0, 0\}$ which represents the fact that all customers are initially free, the c channels are operational free and the orbit is empty.

The firing of the transition $Arrival$ indicates the arrival of a primary call. The service semantics of this transition is ∞-servers (represented by the symbol # placed next to transition) because all free customers are able to generate primary calls. At the arrival of a primary or repeated call to the place $Choice$, if the place Ser_Idle contains at least one operational free channel, the immediate transition $Begin\_Serv$ fires. This firing represents the fact that the customer starts to be served and the server moves into busy state. However, the immediate transition Go_Orbit fires at the arrival of a call who finds no operational free channel i.e. Ser_Idle is empty. Hence, the customer joins immediately the place $Orbit$. Once in orbit, it starts generation of a flow of repeated calls exponentially distributed with rate ν. The firing of transition $Retrial$ represents the arrival of a repeated call from orbit.

When the timed transition $Service$ fires, the customer under service returns to free state (to the place Cus_Free) and the channel becomes idle and ready to serve another customer. The service semantics of transition $Service$ is ∞-servers because several channels can work simultaneously.

A channel can fail during a service period. This is represented by the fact that the transition Act_Fail fires before $Service$ (application of race policy). Thus, the interrupted customer joins the orbit and the failed channel joins the

 ISBN: 978-1-61208-203-5

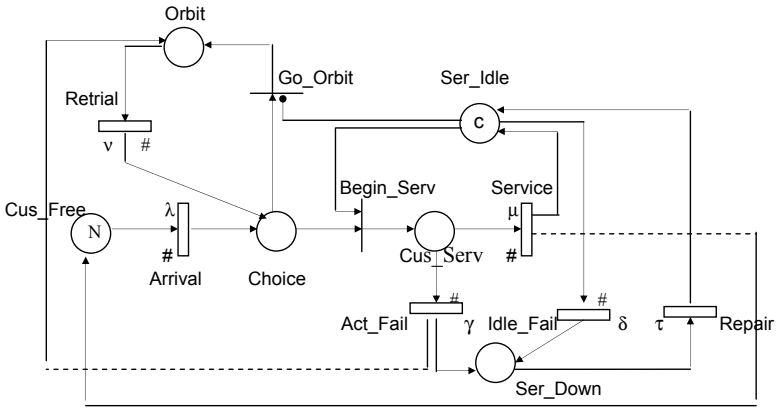


Figure 2. GSPN model of small cell wireless networks with retrials and dependent channels breakdowns

place Ser_Down where it will be immediately repaired. The firing of transition $Repair$ represents the end of the repair time. The repairman repairs one server at a time. Thus, the service semantics of this transition $Repair$ is *single*.

B. Retrial Networks with Dependent Channels Breakdowns

In the model describing dependent breakdowns of channels, depicted in Figure 2, during a service period, a channel can fail, which is represented by the firing of transition Act_Fail. In this case, the interrupted customer joins the orbit and the failed channel joins the place Ser_Down to be repaired. On the other hand, a channel can also fail if it is idle (in the place Ser_Idle). This is represented by the firing of transition $Idle\_Fail$ which has an ∞-servers semantics because several idle channels can also fail in the same time. The transitions Act_Fail and $Idle\_Fail$ representing the breakdown during busy and idle state respectively, may have the same ($\delta = \gamma > 0$) or different rates which correspond to independent and dependent breakdowns disciplines respectively.

C. Retrial Networks with Base Station Breakdowns

In models considering base station breakdowns, all the channels fail down simultaneously and they also return to the operational state at the same time, when the base station is repaired. Hence, the corresponding GSPN models vary slightly from the previous ones. In fact, the models are the same as those given in Figure 1 and Figure 2, in which the multiplicity of the arcs connecting the place Cus_Serv to transition Act_Fail, Act_Fail to the places Ser_Down and $Orbit$, Ser_Down to the transition $Repair$ and the transition $Repair$ to place Ser_Idle equals c (rather that 1), because the c active channels fail down at the same time. The firing of transition Act_Fail will move c tokens in the place Ser_Down, which represents the breakdown of all base station. At the end of the reparation period (after a mean delay of $1/\gamma$), c tokens corresponding to the c channels will be deposited in Ser_Idle. Moreover, in Figure 2, to model the possibility that all channels of the station fail down when being in idle state, we should modify the multiplicity of the arcs connecting the place Ser_Idle to transition $Idle\_Fail$ and $Idle\_Fail$ to place Ser_Down (c rather than 1). The service semantics of the transitions Act_Fail and $Idle\_Fail$ is *single-server semantics*, because the base station is a single unit. Hence, the symbols $\#$ should be omitted from these two transitions.

V. Performance and Reliability Analysis

The aim of this study is twofold. Firstly, we have to verify the correctness of our models and their ergodicity. Next, we derive the formulas of the most important steady-state performance indices.

The primordial qualitative property we have to verify is the *boundness* of the proposed models. This property ensures that each place of the net is bounded and so the model state space is finite. The second important qualitative property is the *liveness*. A transition t is live if from any reachable marking, there is a reachable marking enabling t. Thus, t is live implies that the activity modeled by this transition can always take place from any state. In the proposed models, all transitions are live. Finally, another interesting qualitative property we had to check is the presence of *home states*.

The proposed GSPN models are bounded, live and the initial marking is a home state. Thus, the underlying continuous time Markov chains are ergodic. Hence, the steady-state probability distribution vector π exists and is unique. Once this probability vector is computed, several performance and reliability indices of small cell wireless networks with retrial phenomenon and different breakdowns disciplines can be derived as follows. In these formulas, $M_i(p)$ denotes the number of tokens in place p in marking M_i, A the set of reachable tangible markings, and $A(t)$ is the set of tangible markings reachable by transition t and $E(t)$ is the set of markings where the transition t is enabled.

- `Mean number of busy channels` (n_s): This corresponds to the mean number of tokens in the place Cus_Serv which is also the mean number of customers under service.

$$n_s = \sum_{i:M_i \in A} M_i(Cus\_Serv).\pi_i$$

- `Mean number of customers in orbit` (n_o): This correspond to the mean number of tokens in the place $Orbit$ which models the orbit.

$$n_o = \sum_{i:M_i \in A} M_i(Orbit).\pi_i$$

- `Mean number of operational free channels` (n_d): This represents the average

number of tokens in the place Ser_Idle.

$$n_d = \sum_{i:M_i \in A} M_i(Ser\_Idle).\pi_i$$

- `Mean number of failed channels` (n_f) : This represents the mean number of tokens in the place Ser_Down.

$$n_f = \sum_{i:M_i \in A} M_i(Ser\_Down).\pi_i = s - (n_s + n_d)$$

- `Mean rate of generation of primary calls` $(\overline{\lambda})$: This represents the throughput of the transition $Arrival$.

$$\overline{\lambda} = \sum_{i:M_i \in A(Arrival)} \lambda.M_i(Cus\_Free).\pi_i$$

- `Mean rate of generation of repeated calls` $(\overline{\nu})$: This represents the retrial frequency of customers in orbit. It corresponds to the throughput of the transition $Retrial$.

$$\overline{\nu} = \sum_{i:M_i \in A(Retrial)} \nu.M_i(Orbit).\pi_i$$

- `Mean rate of service` $(\overline{\mu})$: This represents the throughput of the transition $Service$.

$$\overline{\mu} = \sum_{i:M_i \in A(Service)} \mu.M_i(Cus\_Serv).\pi_i$$

- `Mean rate of repair` $(\overline{\tau})$: This represents the throughput of the transition $Repair$.

$$\overline{\tau} = \sum_{i:M_i \in A(Repair)} \tau.M_i(Ser\_Down).\pi_i$$

- `Blocking probability of a primary call` (B_p) :

$$B_p = \frac{\sum_{j:M_j \in A} \sum_{i=1}^{K-s} i.\lambda.Prob[M_j(Cus\_Free) = i \& M_j(Ser\_Idle) = 0]}{\overline{\lambda}}$$

- `Blocking probability of a repeated call` (B_r) :

$$B_r = \frac{\sum_{j:M_j \in A} \sum_{i=1}^{K-s} i.\nu.Prob[M_j(Orbit) = i \& M_j(Ser\_Idle) = 0]}{\overline{\nu}}$$

- `Blocking probability` (B) :

$$B = B_p + B_r$$

- `Utilization of` s `channels` (U_s) : $(1 \leq s \leq c)$ This corresponds to the probability that s servers are busy :

$$U_s = \sum_{i:M_i(Cus\_Serv) \geq s} \pi_i$$

- `Availability of` s `channels` (A_s) : $(1 \leq s \leq c)$ This corresponds to the probability that s servers are operational and idle.

$$A_s = \sum_{i:M_i(Ser\_Idle) \geq s} \pi_i$$

- `Failure probability of` s `channels` (F_s) : $(1 \leq s \leq c)$ This corresponds to the probability that s servers are failed:

$$F_s = \sum_{i:M_i(Ser\_Down) \geq s} \pi_i$$

- `Utilization of the repairman` (U_r) : This corresponds to the probability that at least one server is failed:

$$U_r = F_1 = \sum_{i:M_i(Ser\_Down) \geq 1} \pi_i$$

- `Failure frequency of busy channels` $(\overline{\gamma})$: This represents the throughput of the transition $Failure$ (or $Fail\_Act$) for active breakdowns case and dependent breakdowns case respectively.

$$\overline{\gamma} = \begin{cases} \sum_{i:M_i \in A(Failure)} \gamma.M_i(Cus\_Serv).\pi_i, \\ \text{in active breakdowns,} \\ \sum_{i:M_i \in A(Fail\_Act)} \gamma.M_i(Cus\_Serv).\pi_i, \\ \text{in dependent breakdowns.} \end{cases}$$

- `Failure frequency of idle channels` $(\overline{\delta})$: This represents the throughput of the transition $Fail\_Idle$.

$$\overline{\delta} = \sum_{i:M_i \in A(Fail\_Idle)} \delta.M_i(Ser\_Idle).\pi_i$$

- `Mean waiting time` $(\overline{W})$: The mean waiting time $\overline{W}$ of the customers in the steady state, can be easily obtained with the help of Little's formula:

$$\overline{W} = n_o/\overline{\lambda}$$

- `Mean response time` $(\overline{R})$:

$$\overline{R} = (n_o + n_s)/\overline{\lambda}$$

 ISBN: 978-1-61208-203-5

Table I
VALIDATION OF RESULTS IN RELIABLE CASE

	Reliable model	Active breakdowns of servers	Active breakdowns of station	Dependent breakdowns of servers	Dependent breakdowns of station
Population size	20	20	20	20	20
Number of servers	4	4	4	4	4
Primary call generation rate	0.1	0.1	0.1	0.1	0.1
Service rate	1	1	1	1	1
Retrial rate	1.2	1.2	1.2	1.2	1.2
Server's failure rate	-	1e-25	1e-25	1e-25	1e-25
Server's repair rate	-	1e+25	1e+25	1e+25	1e+25
Mean number of busy servers	1.800748	1.800764	1.800764	1.800763	1.800763
Mean number of customers in orbit	0.191771	0.191786	0.191786	0.191785	0.191785
Mean rate of customers arrivals	1.800748	1.800745	1.800745	1.800745	1.800745
Mean response time	1.106495	1.1065036	1.1065036	1.1065031	1.1065031

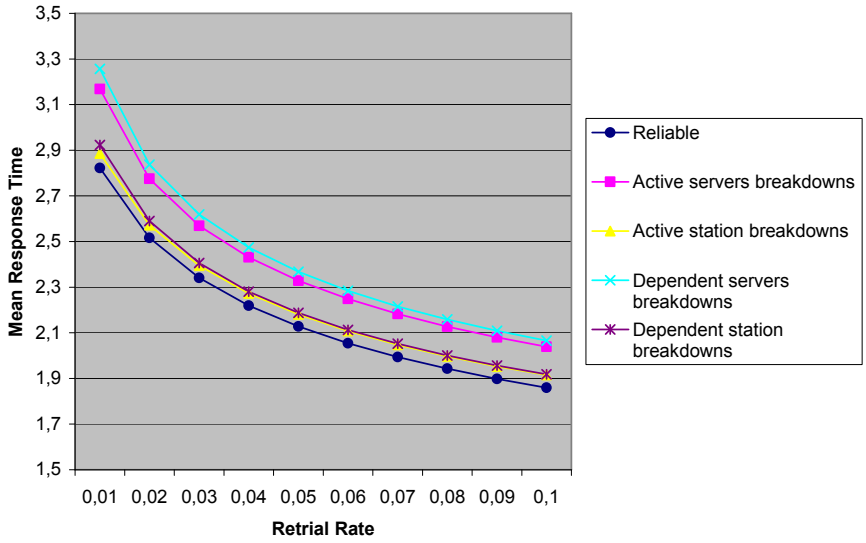


Figure 3. Effect of Retrial Rate on the Mean Response Time with $K = 50$, $c = 4$, $\lambda = 0.4$, $\mu = 5$, $\gamma = 0.02$, $\delta = 0.01$, $\tau = 0.5$

VI. VALIDATION OF MODELS

In this section, we consider some numerical results, to validate the proposed models. To this aim, the results obtained in the reliable case were compared to those obtained by the Pascal program given in the book of Falin and Templeton [4], since if the failure rate in non-reliable models is very low and repair rate is very high, the measures should approach the corresponding ones in reliable models.

In Table 1, we can see that the corresponding performance measures for the model with active channels breakdowns, the model with active station breakdowns, the model with dependent channels breakdowns and the model with dependent station breakdowns are very close to the reliable case. In fact, the derived results are the same up to the 4th decimal digit.

In Figures 3 and 4, the mean response time is splotted versus the retrial rate ν and channels number c respectively. We have presented five curves which correspond to the reliable case, the active and independent channels and station breakdowns disciplines. From Figure 3, we can see how much the increase of retrial rate affects the mean response time which decreases in reliable case and for the different breakdowns disciplines. The numerical results agree with the intuition that the mean response time is better (lower) in the reliable case for all values of the retrial rate. It is also shown from this figure that among the four breakdowns disciplines, the model with active breakdowns of base station gives the best mean response times particularly when the retrial rate is smaller, but when the repeated calls arrive more frequently, the two station breakdowns disciplines (active and dependent) are very close.

In Figure 4, it is demonstrated that the channels number of the base station has a significant influence on the mean response time. We can also see that a small change in

 ISBN: 978-1-61208-203-5

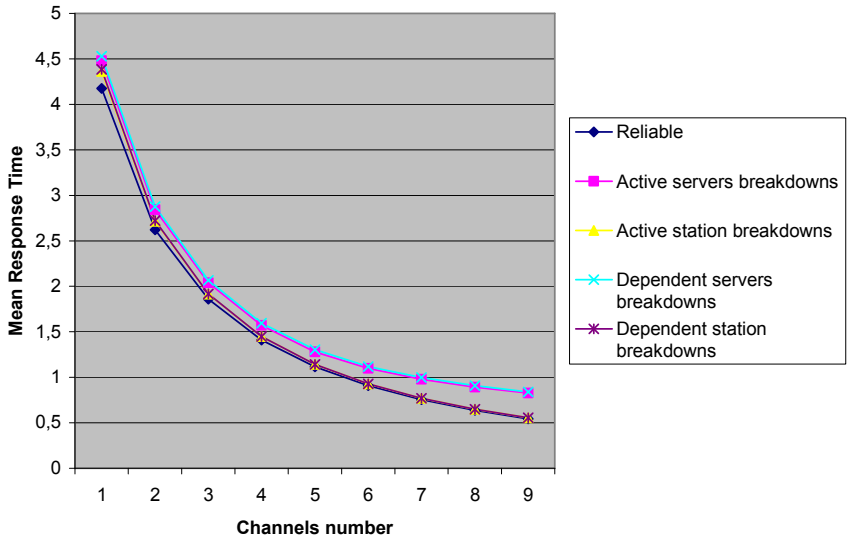


Figure 4. Effect of Channels Number on the Mean Response Time with $K = 50$, $\lambda = 0.4$, $\mu = 5$, $\nu = 1$, $\gamma = 0.02$, $\delta = 0.01$, $\tau = 0.5$

the number of channels, particularly from 1 to 3, produces big difference in the mean response time in reliable and unreliable cases ($\approx -55\%$). However, after a certain value the decrease is not considerable. On the other hand, we can observe that the models with base station breakdowns give the best results, and the worst response time is obtained in dependent breakdowns of channels discipline.

VII. Conclusion

The exponentially increasing demand for wireless data services requires a massive network densification. A promising solution to this problem is the concept of Small Cell Networks, which is founded on the idea of a very dense deployment of short-range, low-cost and low-power base stations.

This paper aims at modeling, performance evaluation and reliability of Small Cell Wireless Networks, taking into account the repeated calls of blocked customers, the finite number of customers served in a cell and the breakdowns of base station channels. Hence, we showed how the behavior of customers in Small Cell Wireless Networks with different breakdowns disciplines can be intuitively described using Generalized Stochastic Petri nets formalism and how several performance and reliability indices can be derived.

References

[1] J. Roszik, J. Sztrik and C. Kim, *Retrial queues in the performance modeling of cellular mobile networks using MOSEL*, Inter. Journal of Simulation: Systems, Science and Technology, vol. 1-2, pp. 38-47, 2005.

[2] J. R. Artalejo and M. J. Lopez-Herrero, *Cellular mobile networks with repeated calls operating in random environment*,Computers & operations research, vol. 37, no7, pp. 1158-1166, 2010.

[3] V. D. Tien, *A new computational algorithm for retrial queues to cellular mobile systems with guard channels*, Computers & Industrial Engineering, vol. 59, pp. 865-872, 2010.

[4] G. I. Falin and J. G. C. Templeton, *Retrial Queues*, Chapman and Hall, London, 1997.

[5] J. R. Artalejo and A. Gómez-Corral, *Retrial Queueing Systems: A Computational Approach*, Springer, Berlin, 2008.

[6] J. R. Artalejo, *Accessible bibliography on retrial queues: Progress in 2000-2009*, Mathematical and Computer Modelling, vol. 51, pp. 1071-1081, 2010.

[7] J. Sztrik and D. Efrosinin, *Tool supported reliability analysis of finite-source retrial queues*, Automation and Remote Control, vol. 71, pp. 1388-1393, 2010.

[8] J. Sztrik and C. S. Kim, *Tool supported performability investigations of heterogeneous finite-source retrial queues*, Annales Univ. Sci. Budapest., Sect. Comp., vol. 32, pp. 201-220, 2010.

[9] J. Wang, J. Cao and Q. Li, *Reliability analysis of the retrial queue with server breakdowns and repairs*, Queueing Systems, vol. 38, pp. 363-380, 2001.

[10] J. Roszik and J. Sztrik, *Performance analysis of finite-source retrial queues with nonreliable heterogenous servers*, Journal of Mathematical Sciences, vol. 146, pp. 6033-6038, 2007.

[11] N. Gharbi and C. Dutheillet, *An algorithmic approach for analysis of finite-source retrial systems with unreliable servers*, Computers and Mathematics with Applications, vol. 62, pp. 2535-2546, 2011.

[12] M. Ajmone Marsan, G. Balbo, G. Conte, S. Donatelli, and G. Franceschinis, *Modelling with Generalized Stochastic Petri Nets*, John Wiley & Sons, New York, 1995.

[13] M. Diaz, *Les réseaux de Petri - Modèles Fondamentaux*, Paris, Hermès Science Publications, 2001.

 ISBN: 978-1-61208-203-5

Theoretically Feasible QoS in a MIMO Cellular Network Compared to the Practical LTE Performance

Mohamed Kadhem Karray
Orange Labs
38/40 rue Général Leclerc, 92794 Issy-Moulineaux France
E-mail: mohamed.karray@orange.com

Miodrag Jovanovic
Orange Labs
38/40 rue Général Leclerc, 92794 Issy-Moulineaux France
E-mail: miodrag.jovanovic@orange.com

***Abstract*—The objective is to build a global analytical approach for the evaluation of the quality of service perceived by the users in wireless cellular networks which is calibrated in some reference cases.**

To do so, a model accounting for interference in a MIMO cellular system is firstly described. An explicit expression of users bit-rates theoretically feasible from the information theory point of view is then deduced. The comparison between these bit-rates and practical LTE performance permits to obtain the progress margins for potential evolution of the technology. Moreover, it leads to an analytical approximate expression of the system performance which is calibrated with the practical one. This expression is the keystone of a global analytical approach for the evaluation of the QoS perceived by the users in the long run of users arrivals and departures in the network. We illustrate our approach by calculating the users QoS as function of the cell radius in different mobility and interference cancellation scenarios.

***Keywords*-MIMO; interference; QoS; cellular; wireless**

I. INTRODUCTION

The performance of a MIMO (*multiple input and multiple output*) cellular network may be considered from different points of view. The information theory gives the ultimate performance of the best possible coding schemes, whereas real systems deploy practical coding schemes of lower performance. On the other hand, information theory gives closed form formulae in several cases, whereas practical system performance is mostly evaluated by simulations such as those of 3GPP (*3rd Generation Partnership Project*) [1].

The objective of the present paper is to compare these two points of views in order to establish an analytical approximation of practical system performance. Our ultimate aim is to build a global analytical approach which is firstly calibrated using simulation results in some reference practical cases, and then used to study the relationships between the key network parameters and the QoS (*quality of service*) perceived by the users.

A. Related work

Telatar [2, Lemma 2] gives the capacity of a MIMO channel with fading and additive white Gaussian noise (without interference) from the information theory point of view. Different MIMO configurations are compared within this context by Foschini and Gans [3] . Blum et al. [4] study the capacity of a MIMO cellular network with flat Rayleigh fading. Tulino and Verdu [5] apply random matrix theory to analyze this capacity. Tarok et al. [6] study the performance of space-time coding which generalizes the Alamouti's codes [7]. Diggavi and Cover [8] study the worst noise process for an additive channel under covariance constraints.

The 3GPP [1] evaluates the performance of LTE systems by simulations. Goldsmith and Chua [9] observed that practical coding schemes performance may be evaluated by a modification of the famous $\log_2(1+\mathrm{SNR})$ Shannon's formula. Mogensen et al. [10] have observed that the LTE capacity in the AWGN context is well approximated by this formula with a multiplicative coefficient. These ideas will be extended in the present paper to MIMO cellular networks with fading.

B. Paper organization

In Section II an explicit expression of users bit-rates feasible from the information theory point of view is given. This expression is compared to practical system performance in Section III. The progress margins for potential evolution of the technology are also presented in this section. Finally, the *global analytical* approach is illustrated in Section IV by evaluating the quality of service perceived by the users in *real* cellular networks accounting for their arrivals and departures.

II. THEORETICALLY FEASIBLE BIT-RATES

The aim of the present section is to establish closed-form expressions of users bit-rates which are theoretically feasible in MIMO cellular networks.

A. Model

Consider a wireless network composed of some disjoint geographic zones, called *cells*, each one being served by a single *base station* (BS) with MIMO antennas. The power transmitted by each BS is limited to some given maximal value. The network operates *Orthogonal Frequency-Division Multiple Access* (OFDMA) which we describe now. The frequency spectrum (allocated to the considered network) is divided into a given number of sub-carriers, which are made available to all base stations. Each BS allocates disjoint subsets of these sub-carriers to its users. Thus, any given user receives only other-BS *interference*; that is the sum of powers emitted by other BS on the sub-carriers allocated to him by his serving BS.

 ISBN: 978-1-61208-203-5

Assume that the bandwidth of each sub-carrier is smaller than the *coherence frequency* of the channel, so that we can consider that the *fading* in each sub-carrier is *flat*. That is, the output of the channel at a given time depends on the input only at the same instant of time. No assumption is made on the *correlation* of the fading processes of the different subcarriers (for a given user and a given BS). However, the fading processes for different users or base stations are assumed independent.

Time is divided into time-slots of length smaller than the *coherence time* of the channel, so that, for a given sub-carrier, the fading remains *constant during each time-slot* and the fading process in different time-slots may be assumed *ergodic*. (Such model for fading generalizes the so-called *quasi-static* model where the fading process at different time-slots is assumed to be independent and identically distributed.)

The codeword duration equals the time-slot, which is assumed sufficiently large so that the *capacity* within each time-slot may be defined in the *asymptotic* sense of information theory. Users perform *single user detection*; thus the interference is added to the *additive white Gaussian noise* (AWGN). The statistical properties of the interference are not known a priory since they depend of the codings of the other users. However the signals transmitted by different base stations are assumed independent.

B. Notations

The *covariance matrix* of a random column vector $X = (X_1, \ldots, X_t)^T$ in $\mathbb{C}^t$ is denoted by $\Gamma_X = E[XX^*]$ where X^* is the transpose complex conjugate of X. Observe for future reference that

$$X^*X = \sum_{j=1}^{t} |X_j|^2 \tag{1}$$

A random vector $(X_1, \ldots, X_t)$ in $\mathbb{C}^t$ is called *circularly symmetric Gaussian* iff it is Gaussian and, each of its components X_i ($i \in [1, t]$) has i.i.d. centred real and imaginary parts.

Consider the downlink of a wireless cellular network. Let u be a base station serving some user through a MIMO channel with t transmitting and r receiving antennas having the following discrete-time model. At a given time instant n the channel output $Y_n \in \mathbb{C}^r$ is related to the channel input $X_{u,n} \in \mathbb{C}^t$ by the following relation

$$Y_n = H_u X_{u,n} + J_n + Z_n, \quad n = 1, 2, \ldots \tag{2}$$

where $H_u \in \mathbb{C}^{r \times t}$ is the fading between the considered user and his serving base station u (fading is assumed constant over time for the moment), $J_n \in \mathbb{C}^r$ is the interference and the random noises $Z_1, Z_2, \ldots$ are i.i.d. with values in $\mathbb{C}^r$ such that each Z_n is circularly-symmetric Gaussian with covariance matrix $\Gamma_{Z_n} = NI_r$ where N is a given positive constant and I_r is the identity matrix of dimension r. The channel input is subject to a power constraint of the form

$$\frac{1}{n} \sum_{k=1}^{n} X_{u,k}^* X_{u,k} \leq P, \quad n = 1, 2, \ldots$$

where P is a given positive constant. Using Equation (1) we see that the above constraint concerns the power aggregated over all the t transmitters.

For each interfering base station $v \neq u$, let $X_{v,n}$ be its transmitted signal and H_v be the fading between the considered user and the base station v (recall that fading is assumed constant over time). Then the interference equals

$$J_n = \sum_{v \neq u} H_v X_{v,n} \tag{3}$$

C. Deterministic fading: Feasible rates

Assume in the present section that the fading is deterministic; i.e. not random.

The *capacity region* of the different users from the information theory point of view (optimized simultaneously over the transmitted signals of all the users) is still unknown. Nevertheless we will show that there is a particular point within this capacity region (i.e. a *feasible* set of users bit-rates) which is easy to establish and express. This point corresponds to the following assumptions:

(A1) The signals transmitted by different base stations are independent.

(A2) The signal $X_{u,n}$ transmitted by the serving base station is circularly-symmetric Gaussian with covariance matrix $\Gamma_{X_{u,n}} = \frac{P}{t} I_t$ (power equi-partition between the transmitting antennas).

(A3) The signal $X_{v,n}$ transmitted by each interfering base station $v \neq u$ is circularly-symmetric Gaussian with a covariance matrix $\Gamma_v = \frac{P}{t} I_t$ (again power equi-partition between the transmitting antennas).

(A4) The transmitted signals are independent from noises.

(A5) The signals transmitted at different time instants are independent.

The covariance matrix of the interference (3) equals

$$\begin{aligned}
\Gamma_J &= E[J_n J_n^*] \\
&= E\left[\sum_{v \neq u} H_v X_{v,n} \sum_{v' \neq u} X_{v',n}^* H_{v'}^*\right] \\
&= \sum_{v \neq u} H_v \Gamma_{X_{v,n}} H_v^* \\
&= \sum_{v \neq u} H_v \Gamma_v H_v^* = \frac{P}{t} \sum_{v \neq u} H_v H_v^*
\end{aligned}$$

where the third equality is due to (A1) and for the fourth one is due to (A3). By (A4), the interference plus noise $Z'_n := J_n + Z_n$ is circularly-symmetric Gaussian [2, Lemma 4] with covariance matrix

$$\mathcal{N} := E[Z'_n Z'^*_n] = NI_r + \frac{P}{t} \sum_{v \neq u} H_v H_v^* \tag{4}$$

Moreover, using (A1)-(A4) the received signal Y_n is circularly-symmetric Gaussian with covariance matrix

$$\Gamma_Y = E\left[(H_u X_{u,n} + Z'_n)\left(X_{u,n}^* H_u^* + Z'^*_n\right)\right] = \frac{P}{t} H_u H_u^* + \mathcal{N}$$

Assumption (A5) permits to restrict ourselves to the mutual information at a given time-instant; that is

$$\begin{aligned}
I(X_{u,n}; Y_n) &= h(Y_n) - h(Y_n | X_{u,n}) \\
&= h(Y_n) - h(Z'_n) \\
&= \log_2 \det(\pi e \Gamma_Y) - \log_2 \det(\pi e \mathcal{N}) \\
&= \log_2 \det\left(I_r + \frac{P}{t} H_u H_u^* \mathcal{N}^{-1}\right)
\end{aligned}$$

 ISBN: 978-1-61208-203-5

where for the first equality we use [11, Theorem 1.6.2], for the third one we use [2, Lemma 2] and where $\mathcal{N}$ is given by (4). Recall that the capacity from the information theory point of view, denoted by C, is the supremum of the mutual information over all the distributions of the input signal. Thus

$$C \geq \log_2 \det \left(I_r + \frac{P}{t} H_u H_u^* \mathcal{N}^{-1} \right)$$

The right-hand side of the above equation[1] gives a *feasible* bit-rate for the considered user. Since our assumptions (A1)-(A5) are the same for all the users, we get similar expressions for the feasible bit-rates of the other users and this *collection of bit-rates* of the different users is *feasible*.

Till now we didn't account for the propagation-losses L_u and $\{L_v\}_{v \neq u}$ induced by the distance and shadowing between the considered user and the serving and interfering base stations respectively. In order to account for these losses, the above formula should be modified as follows

$$C \geq \log_2 \det \left(I_r + \frac{P}{t} \frac{H_u H_u^*}{L_u} \mathcal{N}^{-1} \right) \quad (5)$$

where the noise plus interference covariance matrix $\mathcal{N}$ is now given by

$$\mathcal{N} = N I_r + \frac{P}{t} \sum_{v \neq u} \frac{H_v H_v^*}{L_v} \quad (6)$$

Remark 1: Continuous-time. Consider a continuous-time model of the channel. Let w be the bandwidth of the considered sub-carrier. The results in the discrete-time extend to the continuous-time case, but the capacity bounds, such as the right-hand side of (5), should be multiplied by the bandwidth w of the considered sub-carrier. In other words, the $\log_2(\cdot)$ should be replaced by $w \times \log_2(\cdot)$.

D. Ergodic capacity

Consider now a given sub-carrier and multiple time-slots. Recall that we assumed that the fading for different time-slots are independent and identically distributed. By the law of large numbers, the capacity averaged over a large number of time-slots would approach the so-called *ergodic capacity* $E[C]$ where the expectation is with respect to the fading states. No assumption is made on the distribution of the fading matrix H_v except that its covariance equals identity; that is

$$E[H_v H_v^*] = I_r, \quad \text{for all BS } v$$

which means that the fadings of two different transmitting antennas are decorrelated and that the fading second moment for a given antenna equals 1. The following proposition gives a lower bound for the ergodic capacity.

Proposition 1: The ergodic capacity of the channel (2) is lower bounded by

$$E[C] \geq E[\log_2 \det (I_r + \mathrm{SINR} H_u H_u^*)] \quad (7)$$

where the expectation is with respect to the fading H_u with the serving BS and

$$\mathrm{SINR} = \frac{(P/t)/L_u}{N + (P/t) \sum_{v \neq u} 1/L_v} \quad (8)$$

which may be viewed as the *Signal to Interference and Noise Ratio* per transmitting antenna[2].

[1] which is coherent with [4, Equation (2)]

[2] See [5, Equation (3.169)].

Proof: Note that the expectations in the present proof are with respect to the fading random matrices with the serving BS H_u and with the interfering BS $\{H_v\}_{v \neq u}$. Let $E[\cdot|H_u]$ designates the expectation conditionally to H_u. By the properties of the conditional expectation, we have $E[C] = E[E[C|H_u]]$. Equation (5) implies that

$$E[C|H_u] \geq E\left[\log_2 \det \left(I_r + \frac{P}{t} \frac{H_u H_u^*}{L_u} \mathcal{N}^{-1} \right) \middle| H_u \right]$$

where $\mathcal{N}$ is given by (6). Using the convexity of the function $\mathcal{N} \mapsto \log_2 \left[\det \left(I_r + \frac{P}{t} \frac{H_u H_u^*}{L_u} \mathcal{N}^{-1} \right) \right]$ on the set of positive definite matrices of $\mathbb{C}^{rxr}$ (see [8, Lemma II.3]) and Jensen's inequality, we deduce that

$$\begin{aligned} E[C|H_u] &\geq \log_2 \det \left(I_r + \frac{P}{t} \frac{H_u H_u^*}{L_u} E[\mathcal{N}|H_u]^{-1} \right) \\ &= \log_2 \det \left(I_r + \frac{P}{t} \frac{H_u H_u^*}{L_u} E[\mathcal{N}]^{-1} \right) \\ &\geq \log_2 \det (I_r + H_u H_u^* \mathrm{SINR}) \end{aligned}$$

where SINR is given by (8). Thus

$$E[C] = E[E[C|H_u]] \geq E[\log_2 (1 + H_u H_u^* \mathrm{SINR})]$$

■

The right-hand side of (7) may be calculated by using [2, Theorem 2].

III. Theoretical versus practical performance

The objective of the present section is to compare the theoretical expression established in the previous section to practical LTE performance.

A. AWGN

Consider firstly a user served by a base station through an additive white Gaussian noise (AWGN) SISO channel without neither fading nor interference for the moment. The user gets ideally (i.e. in the asymptotic sense of information theory) a bit-rate given by the famous Shannon's formula $w \log_2 \left(1 + \frac{P/L_u}{N} \right)$ where w is the bandwidth allocated to the considered user, N is the noise power, P is the power transmitted by the BS and L_u is the propagation-loss (thus P/L_u is the received power). In order to get rid of the dependence of the bit-rate on the bandwidth, we define the *spectral efficiency* as the ratio of the bit-rate by the bandwidth which equals $\log_2 \left(1 + \frac{P/L_u}{N} \right)$ in the AWGN context.

Mogensen et al. [10] and the 3GPP [12, §A.2] have observed that the LTE system spectral efficiency in this AWGN context is well approximated by

$$s \simeq a \log_2 \left(1 + \frac{P/L_u}{N} \right) \quad (9)$$

for some constant $a < 1$ accounting on the one hand for the gap between the practical coding schemes and the optimal ones and on the other hand for the loss of capacity due to signalling. This observation shall be confirmed and the typical value of a for LTE will be given. Note that the relative difference $1 - a$ between the Shannon's limit and the practical LTE system may be seen as a progress margin for potential evolution of the technology in the AWGN context.

In the AWGN context, the 3GPP [12, §A.2] shows that there is a 25% gap between the practical coding schemes and

 ISBN: 978-1-61208-203-5

the optimal ones (in the sense information theory). Moreover, some of the transmitted bits are used for signalling which induces a supplementary capacity loss of about 30% (see [13, §6.8],[14, p.155]); this leads to $a = 0.75 \times (1 - 0.3) \simeq 0.5$ in Equation (9). Figure 1 shows that the spectral efficiency obtained by simulations with Orange's link tool *agrees* with the analytic approximation (9) with $a = 0.5$.

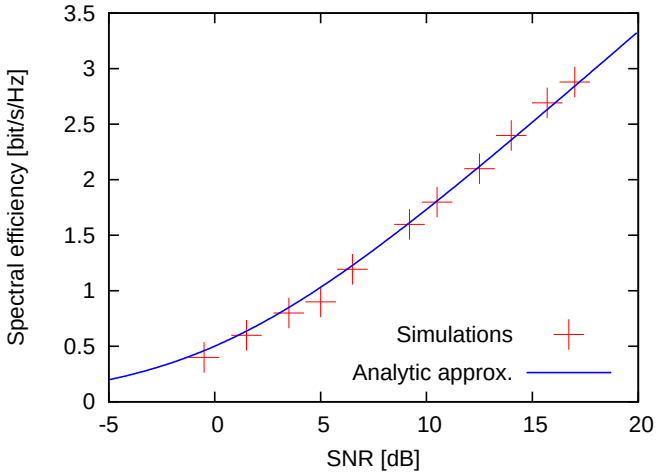


Fig. 1. Practical performance for AWGN

Remark 2: Indeed the signalling loss depends on the number of transmitting and receiving antennas. It is about $40/168 = 24\%$, $48/168 = 29\%$ and $52/168 = 31\%$ respectively for SIMO 1×2, MIMO 2×2 and MIMO 4×2 (see [13, §6.8],[14, p.155]).

B. Fading and interference

We aim now to account for fading, MIMO and interference. In this context, let the *spectral efficiency* be the ratio of the bit-rate averaged over the fading (called ergodic capacity in the information theory framework) by the bandwidth.

In order to simplify the notation, we denote by S the analytical (lower bound of the) spectral efficiency given in the right-hand side of (7) pondered by the parameter $a = 0.5$ obtained in the previous section; that is

$$S\left(\mathrm{SINR}, t, r\right) = aE\left[\log_2 \det\left(I_r + H_u H_u^* \mathrm{SINR}\right)\right] \qquad (10)$$

where SINR is the signal to interference and noise ratio (per transmitting antenna) given by Equation (8).

The question now is what is the practical LTE spectral efficiency compared to the above analytical expression? Is it better or worse and what is the difference?

In order to get the practical LTE performance, we consider the output of Orange's simulator compliant with the 3GPP recommendation [1] (see this reference for the details of the simulations) in the so-called *calibration* case. It corresponds to MIMO 1×2 with *round robin* (RR) scheduler. We consider also other MIMO configurations and *proportional fair* (PF) scheduler, keeping all the other parameters unchanged. In particular, each base station always transmits its maximal power (*full buffer*).

According to [1], the 3GPP simulations consist in generating several realizations of the users positions, shadowing losses and fading channels. For each user location and each shadowing realization, the spectral efficiency is averaged over a large number of fadings samples (about 1000). The value of the SINR including only the distance and the shadowing effects (and not fading) is also given. Then the spectral efficiency as function of the SINR is compared to the theoretical relation (10). More specifically, we make a linear regression between the spectral efficiency obtained from simulations and the theoretical efficiency given by Equation (10); that is we search for some b such that

$$s \simeq b \times S\left(\mathrm{SINR}, t, r\right) \qquad (11)$$

MIMO	Scheduler	b	residual stand. dev.	b'
1×2	RR	0.83	0.45	0.98
1×2	PF	1.02	0.65	1.19
2×2	PF	0.67	0.74	1.08
4×2	PF	0.49	0.76	0.90

TABLE I
RESULTS OF THE LINEAR FITTINGS.

Table I gives the results of the linear fitting (11); i.e. the values of b and the corresponding residual standard deviation for different MIMO configurations[3] (the first row corresponds to the calibration case [1, Table A.2.2-1]). Moreover, the 95%-confidence interval is about $b \pm 0.01$ for all the studied cases.

Figure 2 shows the spectral efficiency as function of the SINR from simulations and from the analytical expression (right-hand side of (11)) for the calibration case. Observe that the analytical expression reproduces well the general tendency of the empirical data obtained from simulations. The figures for the other cases listed in Table I are also generated, but not reproduced in the paper due to their similarity to Figure 2.

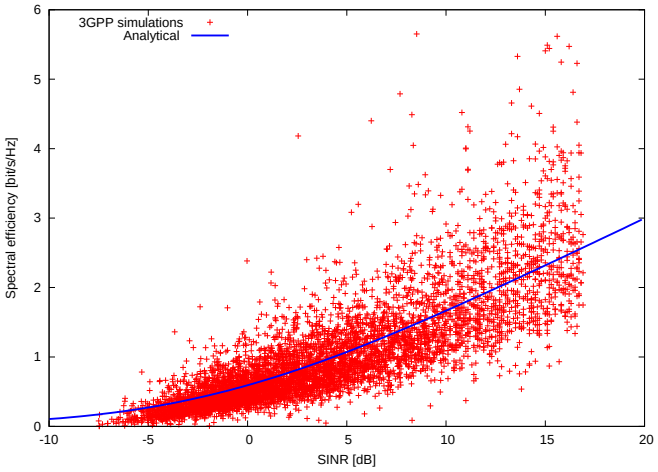


Fig. 2. Simulations versus the analytical expression (right-hand side of (11)) for the calibration case

Remark 3: In order to simplify the calculations we have also tested a linear regression between the spectral efficiency s obtained from simulations and the AWGN expression (9).

[3]All the considered cases have a MRC (Maximum Ratio Combining) receiver, except the MIMO 4×2 case which has a MMSE (Minimum Mean Square Error) receiver. At the base station side, the transmitting antennas are pairwise cross-polar. In the case MIMO 4×2, the two cross-polar pairs of transmitting antennas are separated by 10 times the wavelength.

 ISBN: 978-1-61208-203-5

Observe from Equation (8) that when noise is dominant against interference, then

$$\mathrm{SINR} = \frac{(P/t)/L_u}{N} = \frac{P/L_u}{N} \times \frac{1}{t}$$

Thus, in this particular case, the term $\frac{P/L_u}{N}$ in the right-hand side of (9) equals $\mathrm{SINR} \times t$. Then, in the general case, it is natural to look for a fitting in the form

$$s \simeq b' \times a \log_2 (1 + \mathrm{SINR} \times t)$$

The resulting values of b' are indicated in Table I with residual standard deviations close to those indicated in the fourth column of that table.

C. SINR

For the analytical approach we use a similar geometric pattern of the network (hexagonal) and the same propagation-loss modeling regarding the distance and shadowing effects (fading has been already taken into account on the link level in the previous section) as the 3GPP calibration case [15, Table A.2.1.1-3] and [1, Table A.2.2-1].

More specifically, the frequency carrier is 2GHz. The distance loss model is $L = 128.1 + 37.6 \times \log_{10}(r)$ [in dB]. A supplementary penetration loss of 20dB is added. The shadowing is modeled as a centered log-normal random variable of standard deviation 8dB. The following 2D horizontal antenna pattern is used

$$A(\varphi) = -\min\left(12\left(\frac{\varphi}{\theta}\right)^2, A_\mathrm{m}\right), \quad \theta = 70^\circ, A_\mathrm{m} = 20\mathrm{dB} \tag{12}$$

The system bandwidth is $W = 10$MHz, the noise power equals $N = -95$dBm (-174dBm/Hz, noise figure=9dB) and the transmission power of the base station is $P = 60$dBm (46dBm plus $G = 14$dBi of antenna gain). The network is composed of 36 hexagons (6×6). Each hexagon comprises three sectors which gives a total of 108 sectors. The distance between the centers of two neighboring hexagons is 500m. We generate 3600 random user locations uniformly in the network; that is 100 user locations per hexagon in average.

The 3GPP simulations published in [1] are made on a planar network with random locations of the users. In the present study, two network models are considered: either planar or toroidal (to avoid the border effects).

Each mobile is served by the base station with the smallest *propagation-loss* (including distance, shadowing and antenna pattern). In order to facilitate the comparison of our results to those of 3GPP, we define the coupling-gain as the antenna gain G minus propagation-loss L with the serving base station. The cumulative distribution function (CDF)[4] of the coupling-gain obtained by 3GPP simulations [1, Figures A.2.2-1 (left)] and by our models are given in Figure 3. This figure shows that the results of our planar network are close to those of 3GPP simulations, whereas those of toroidal network give larger coupling gain. This is due to the fact that in a planar network edge users get smaller coupling gain than in the toroidal one.

The SINR for each mobile is calculated by Equation (8), where u is the index of the serving base station. Figure 4 shows the CDF of the SINR coming from 3GPP simulations [1, Figure A.2.2-1 (right)] compared to that resulting from our models. Again our planar model gives closer results to the 3GPP simulations than the toroidal one. Nevertheless, the difference between the SINRs of the toroidal and the planar networks is smaller than 0.5dB.

[4]over all the user locations in the network

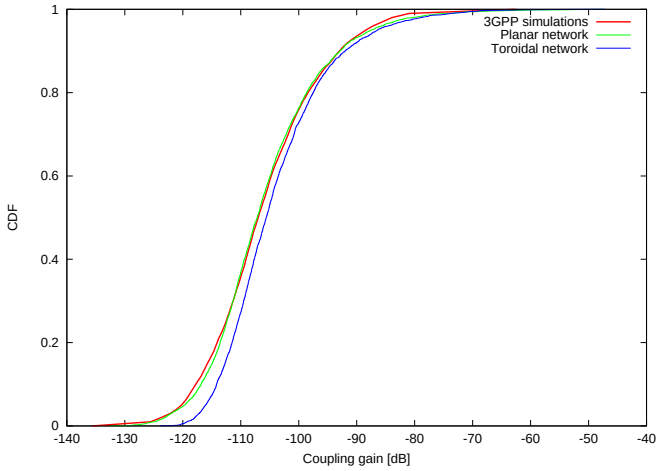


Fig. 3. CDF of the coupling gain (antenna gain minus propagation loss)

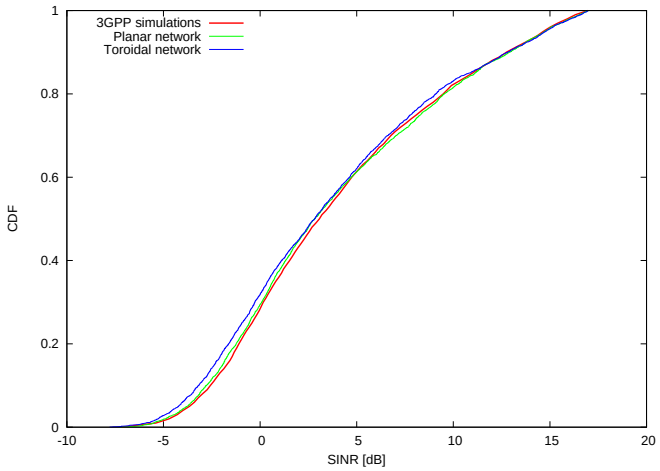


Fig. 4. CDF of SINR

Remark 4: Figure 4 shows that the SINR doesn't exceed 17dB. Indeed, each mobile served by a given base station (sector) is at least interfered by the two other sectors on the same site. The power received from each of these sectors is at least 10^{-2} times that received from the serving BS (this is related to $A_\mathrm{m} = 20$dB in Equation (12)). The interference to signal ratio is consequently larger than 2×10^{-2} i.e. -17dB which explains the observed upper limit of SINR.

Remark 5: Observe that the SINR defined by Equation (8) is different from the SINR calculated by 3GPP simulations which equals

$$\mathrm{SINR}_{3\mathrm{GPP}} = \frac{P/L_u}{N + P\sum_{v \neq u} 1/L_v}$$

However, if noise is negligible compared to interference, then the two SINRs are identical. This is the case in the considered urban scenario (small cell radius), so we have not to distinguish between these two SINRs.

 ISBN: 978-1-61208-203-5

		Arithmetic mean		Harmonic mean	
MIMO	Scheduler	Simus	Analytic	Simus	Analytic
1×2	RR	1.01	1.00	0.50	0.69
1×2	PF	1.32	1.23	0.80	0.85
2×2	PF	1.43	1.41	0.84	1.00
4×2	PF	1.54	1.54	0.95	1.18

TABLE II
CELL SPECTRAL EFFICIENCY: COMPARISON OF THE 3GPP SIMULATIONS AND THE ANALYTIC RESULTS.

D. Spectral efficiency

For each mobile we calculate the spectral efficiency corresponding to its SINR by relation (11). In order to facilitate the comparison of our results to those of 3GPP, we define the normalized user throughput as the spectral efficiency divided by 10 (this is historically related to the fact there are 10 users per cell in 3GPP simulations). The CDFs of the normalized user throughput obtained by 3GPP simulations [1, Figure A.2.2-3 (left)] and by our model are plotted in Figure 5. The 3GPP distribution is more spread than that of our models; this is related to the fact that the 3GPP spectral efficiency represents some variability around the analytic one as shown in Figure 2. Moreover, we observe that the results of the planar and toroidal models for the network are close to each other. Thus, the toroidal model is considered for the remaining part of the paper.

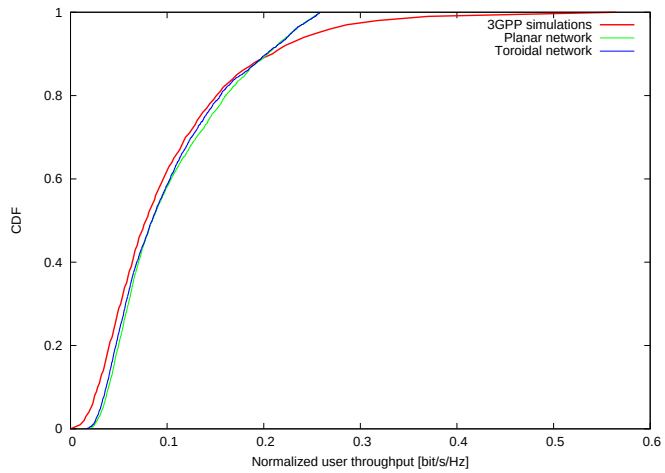


Fig. 5. CDF of normalized user throughput

Table II gives the *arithmetic mean* of the spectral efficiencies at the different locations (called cell spectral efficiency) for both 3GPP simulations and analytic approach. The results of two methods agree for all the considered MIMO and scheduler configurations.

Remark 6: Note that the results of the simulations given in Table II are produced by the simulator of Orange which is one of the contributors to 3GPP. The values indicated in [1, Table A.2.2-2] are in fact averaged over the different 3GPP contributors including Orange. In particular, for the calibration case (MIMO 1×2 with RR scheduler) Orange's result is 1.01 whereas 3GPP average is 1.1. The variability of the results amoung the contributors is partially due to the randomness induced by the shadowing.

IV. USER'S QOS CALCULATION

In order to illustrate the whole analytical approach, we show now how to calculate the QoS perceived by the users in a dynamic context; i.e., when users arrive and depart from the network. Variable bit-rate (VBR) calls such as mail, http, ftp are considered. Each VBR call aims to transmit some volume of data at a bit-rate which is decided by the network. Define the *peak bit-rate* at a given location as the bit-rate which may be allocated to some user in this location assuming that he is alone in the cell and that all the base stations transmit at their maximal powers. As observed by Caire and al. in [16, §I], for the VBR calls the performance at the link level for a given location should be firstly averaged over the fading; then these averages may be used at the queueing theory time scale to account for call arrivals and departures. Therefore we have a natural separation of the time scales of information theory and queueing theory. Assuming a round robin schedular, the peak bit-rate at each location equals the system bandwidth times the spectral efficiency at that location given by Equation (11).

Let ρ be the traffic demand (in bit/s) per cell; that is the ratio of the average volume of data per call to the duration between two call arrivals to the cell. Assume that the traffic demand is uniformly distributed over the cell and that the users are allocated equal portions of the available resources (time and/or frequency). We assume that the users don't move during their calls.

In this context, queueing theory [17], [18, Example 10] shows that the user's throughput in the long run of the call arrivals and departures is given by

$$\bar{r} = \max(0, \rho_c - \rho) \tag{13}$$

where ρ_c is the so-called *critical traffic demand* which equals the *harmonic mean* of the peak bit-rates in the cell if the users don't move during their calls. On the opposite, if the users move during their calls, then the *critical traffic demand* equals the *arithmetic mean* of the peak bit-rates in the cell. At high mobility, the above formula holds also true with the appropriate critical traffic demand.

Consider the numerical setting of the calibration case described in Section III-C. Figure 6 shows the throughput per user in the cell as function of the cell radius for a traffic demand density 300kbit/s/km^2 (typical value in urban areas) for both the no-mobility and high mobility cases. We consider also the case when the interference is completely cancelled. As expected, the user's throughput decreases with the cell radius and ultimately vanishes for some critical cell radius. Moreover, observe that the mobility improves the user throughput from the queueing theory point of view as proved theoretically in [19, §VI]. On the other hand, observe that the interference cancellation improves performance, but this improvement decreases as the cell radius increases. For a cell radius of 0.3km, the interference cancellation increases the user throughput by a factor of about 7; whereas this factor is only of about 2 for cell radius of 2km. This is due to the fact that as the cell radius inceases, the noise becomes dominant compared to interference.

The analytical approach developed in the present paper permits to make the parametric study represented in Figure 6

 ISBN: 978-1-61208-203-5

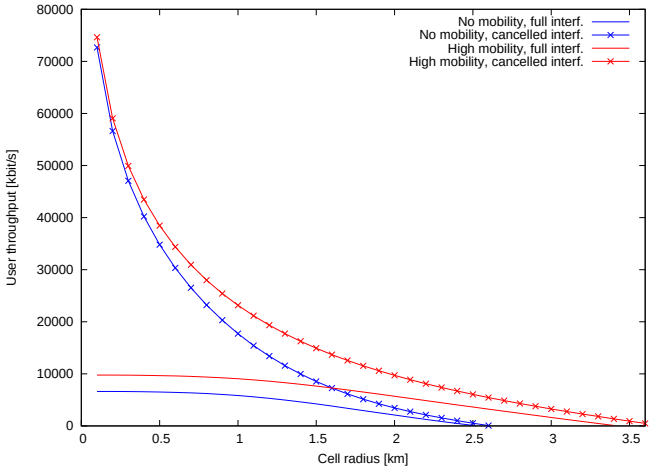


Fig. 6. User throughput as function of the cell radius in the cases of no mobility and high mobility (for the calibration scenario)

in about one minute; whereas it would require weeks for the 3GPP simulations.

Remark 7: Table II shows the harmonic means of the spectral efficiency obtained from 3GPP simulations and from the analytical expression. The difference may be explained as follows. Recall that the harmonic mean is sensitive to the minimal value of the considered data; for example if one of these data is null then the harmonic mean vanishes. Moreover, Figure 2 shows that the 3GPP spectral efficiency represents some variability around (and in particular comprise smaller values than) the analytic curve. This explains why the harmonic means obtained from simulations in Table II are lower than the analytic ones.

V. Conclusion

We describe a simple model of a MIMO cellular network which permits to obtain an analytical expression of users bit-rates which are feasible from the information theory point of view. This expression accounts for the variety of MIMO configurations (numbers of transmitting and receiving antennas) and radio conditions (SINR). This expression is compared to practical LTE performance evaluated by 3GPP simulations for different cases including the so-called calibration case. The comparison shows that the analytical expression may be adjusted to the practical performance by a multiplicative coefficient which depends on the MIMO configuration but not on the SINR. Additionally, we show the progress margin for potential evolution of the technology.

In order to illustrate the whole analytical approach, we calculate the throughput perceived by the users in the long run of users arrivals and departures in the network. The analytical approach permits to make the calculations in a much faster computing time than a purely simulation approach. The comparison of null and high user's mobility permits to quantify the effect of this mobility from the queueing theory point of view. Studying the case when inteference is completely cancelled permits to quantify the ultimate impovement expected from the interference cancellation.

The simplifying assumption that the base stations are always transmitting (even when there are no users to serve) shall be examined in the future work. Moreover, the QoS for other service classes such as streaming will be studied.

Acknowledgement: We thank B. Błaszczyszyn (INRIA), M. Debbah (Supelec) as well as A. Saadani, S. Jeux and A. Jassal (Orange Labs) for useful discussions related to the present paper.

References

[1] 3GPP, "TR 36.814-V900 Further advancements for E-UTRA - Physical Layer Aspects," in *3GPP Ftp Server*, 2010.

[2] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *AT&T Technical Memorandum*, June 1995.

[3] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, March 1998.

[4] R. S. Blum, J. H. Winters, and N. R. Sollenberger, "On the capacity of cellular systems with MIMO," *Communications Letters, IEEE*, vol. 6, no. 6, pp. 242 –244, jun 2002.

[5] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 1, 2004.

[6] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block coding for wireless communications: performance results," *IEEE J. Select. Areas Commun.*, vol. 17, no. 3, pp. 451–460, 1999.

[7] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Select. Areas Commun.*, vol. 16, no. 8, pp. 1451 –1458, oct 1998.

[8] S. Diggavi and T. Cover, "The worst additive noise under a covariance constraint," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 3072–3081, 2001.

[9] A. J. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218–1230, 1997.

[10] P. E. Mogensen, W. Na, I. Z. Kovács, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. E. Kolding, K. Hugl, and M. Kuusela, "LTE Capacity Compared to the Shannon Bound," in *Proc. of VTC Spring*, 2007, pp. 1234–1238.

[11] S. Ihara, *Information theory for continuous systems*. World Scientific, 1993.

[12] 3GPP, "TR 36.942-V830 Evolved Universal Terrestrial Radio Access (E-UTRA) - Radio Frequency (RF) system scenarios," in *3GPP Ftp Server*, Sep. 2010.

[13] ——, "TR 36.211-V910 Evolved Universal Terrestrial Radio Access (E-UTRA) - Physical Channels and Modulation," in *3GPP Ftp Server*, Mar. 2010.

[14] E. Dahlman, S. Parkvall, and J. Skold, *4g: LTE/LTE-advanced for mobile broadband*. Academic Press Inc, 2011.

[15] 3GPP, "TR 25.814-V710 Physical layer aspects for evolved Universal Terrestrial Radio Access (UTRA)," in *3GPP Ftp Server*, 2006.

[16] G. Caire, G. Taricco, and E. Biglieri, "Optimum power control over fading channels," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1468–1489, Jul. 1999.

[17] J. W. Cohen, *On regenerative processes in queueing theory*, ser. Lecture Notes in Economics and Mathematical Systems. Springer Berlin Heidelberg, 1976, vol. 121.

[18] M. K. Karray, "User's mobility effect on the performance of wireless cellular networks serving elastic traffic," *Wireless Networks (Springer)*, vol. 17, no. 1, Jan. 2011.

[19] T. Bonald, S. Borst, and A. Proutière, "How mobility impacts the flow-level performance of wireless data systems," in *Proc. of IEEE INFOCOM*, 2004, pp. 1872–1881.

 ISBN: 978-1-61208-203-5

Evaluation of Routing Protocols for Internet-Enabled Wireless Sensor Networks

Ion Emilian Radoi
School of Informatics
The University of Edinburgh
Edinburgh, United Kingdom
e-mail: emilian.radoi@gmail.com

Aditi Shenoy
School of Informatics
The University of Edinburgh
Edinburgh, United Kingdom
e-mail: shenoy.aditi@gmail.com

D. K. Arvind
School of Informatics
The University of Edinburgh
Edinburgh, United Kingdom
e-mail: dka@inf.ed.ac.uk

Abstract - **This paper investigates the choice of routing algorithms for a CoAP-UDP stack for a internet-enabled Wireless Sensor Network (WSN) running an application for emergency monitoring and evacuation of people in a building. The routing protocols considered belong to two classes: proactive protocols (CTP, RPL) and reactive protocols (AODV, DSR). The emergency monitoring and evacuation scenario, running on a WSN with a full stack, was modelled and simulated in the SpeckSim behavioural simulator. The results of our study demonstrated that AODV would be the protocol of choice for the chosen application. The methodology advocated is sufficiently general for investigating protocol choices for other applications.**

Keywords – WSN; routing protocols; CoAP; AODV; RPL.

I. INTRODUCTION

Internet-enabled WSNs can be used to bridge the physical world that we inhabit with the virtual world of the Internet. Miniature battery-operated sensors with wireless connectivity and processing capability which are attached to objects can be used to extend the connectivity of the Internet. Information from the sensory data can be used to build web-oriented applications such as smart metering and smart building networks, and a number of bodies have been active in their standardisation.

The Internet Protocol for Smart Objects (IPSO) Alliance [17] has been involved in the interfacing of IP technology with everyday physical devices. In addition, the Internet Engineering Task Force (IETF) has incorporated several Working Groups towards the standardization of IP protocols for these objects. Their first attempt was to compress IPv6 over Low power Wireless Personal Area Networks (6LoWPAN) [1] to enable its use in low-power 802.15.4 radios. The Routing Over Low power and Lossy networks (ROLL) Working Group is promoting a routing protocol called the IPv6 Routing Protocol for Low-power and Lossy networks (RPL) [2].

There is now a progress from concern about network connectivity between physical objects (actuators, sensors, embedded devices) and the Internet, towards building useful web service-oriented applications over this basic layer of connectivity.

Internet-enabled WSNs can be realised by adapting traditional web protocols in ways suitable to different applications, thereby enabling the integration of these sensor-enriched physical objects to the Internet. This can be made possible if the existing REpresentational State Transfer (REST) architectural style can be extended to accommodate new application layer protocols suitable for WSNs over existing transport protocols such as TCP/ UDP.

The IETF Constrained RESTful Environments (CoRE) Working Group [18] is focusing on designing application layer protocols that manipulate sensor data, which overcome the restrictions of their networking environments. The resulting Constrained Application Protocol (CoAP) [3] integrates the different facets of the web service architecture. CoAP includes a subset of the REST features that are available in HTTP, to enable effective Machine-to-Machine (M2M) communication between devices.

The question asked in this study was that for a given choice of application/transport layer protocol (CoAP/UDP), and a data link layer protocol (SpeckMAC-D [16]), what is the appropriate choice of routing protocol for the given application scenario of emergency monitoring and evacuation of people in a building.

We considered two sets of routing protocols classified on the basis of their gathering and maintenance of routing information. Proactive protocols generate routing tables and periodically exchange update information, and reactive ones which do not, but instead trigger a discovery process when routing information is required. We selected RPL and Collection Tree Protocol (CTP) from the proactive class, and Ad hoc On-Demand Distance Vector Routing (AODV) and Dynamic Source Routing (DSR) from the reactive one.

We have implemented CoAP-UDP over each of the chosen routing protocols, and have examined in each case the behaviour of the resulting protocol stack. Based on a selection of evaluation metrics relevant to constrained networks, we determine the suitability of the routing protocols for ensuring effective, reliable communication between resource-constrained devices.

Section II reviews related work in this field; Section III describes the different protocols that were implemented; Section IV describes the emergency monitoring and evacuation application and its implementation in the simulator, along with the implementation of the routing protocols. Section V provides an analysis of the results and Section VI presents the concluding remarks.

II. RELATED WORK

The challenge in achieving WSN interoperability with IP networks has been recognised [10], and so has the need for an open resource-oriented architecture for building web services in sensor networks.

A few research papers have concentrated on the need for a new application protocol such as CoAP, and justify its use

 ISBN: 978-1-61208-203-5

in WSNs. Colitti et al. [4] provide a detailed description of CoAP and compare it with HTTP by running the Contiki and libcoap [20] CoAP versions. They demonstrated that the energy consumption of CoAP-running sensor nodes is significantly lower than those running HTTP. Kovatsch et al. [19] describe an implementation of the IETF CoAP protocol for the Contiki operating system that leverages the ContikiMAC low-power duty cycling mechanism to provide power efficiency.

Kuladinithi et al. [5] describe CoAP's Contiki and TinyOS implementations to integrate CoAP into an existing WSN-oriented logistics system for cargo containers. The focus is on performing CoAP-HTTP comparisons based on certain application-specific evaluation metrics such as data retrieval and access rates/times. In this paper we evaluate the entire protocol stack, based on five evaluation metrics.

Both simulated and real implementations of RPL have been evaluated [6][7] since the protocol was selected as the IETF candidate for standardization in WSNs.

Several papers and dissertations [15] have compared different routing protocols for WSNs, but few comparisons have been made between the protocols that we have considered in this study. In [11], for instance, on-demand routing protocols such as AODV, DSDV and DSR are evaluated using the NS2 simulator which concluded that AODV outperforms the other two protocols in terms of packet delivery ratio. In [14], the same protocols were evaluated with similar results in terms of packet delivery, but higher performance was demonstrated by DSDV as the network was scaled and a radio shadowing model was considered.

A number of simulators have been developed for understanding the behaviour of WSNs [10]. The Cooja simulator [11], for example, is focussed on simulating hardware details of the WSN nodes. TOSSIM [12] is a discrete event application-level simulator that can be used for TinyOS-based WSNs. The former is better suited for analysing the impact of low-level network details at cycle-level accuracies, whereas the latter is better suited for capturing the impact of application-specific issues on performance.

In this study, we used the SpeckSim [13] behavioural-level simulation environment which has been designed to perform evaluation across the different layers of a protocol stack to determine the most efficient set of protocols for a given class of applications.

III. BACKGROUND

This section briefly presents the protocols that were chosen for this study. It is followed by the implementation section which further discusses them in more detail.

A. The Constrained Application Protocol (CoAP)

The interfacing of resource-constrained embedded devices to the Internet requires extensions to its current architecture and new light-weight representations. HTTP is less able to handle M2M interactions efficiently with the additional overhead of heavy-weight resource representation formats such as HTML and XML. There is a need for a compact REST-affiliated architectural style to connect internet-enabled physical objects and access them through universally accepted standards-based methods.

CoAP is a generic web protocol, defined by the IETF CoRE Working Group [3], which aims to enable interoperability between embedded constrained M2M applications. The goal of this protocol is not only to compress HTTP, but to include constraints such as statelessness, cache-ability, layered system, uniform interface common in current web protocols and additional features such as multicast support, built-in device discovery, asynchronous message exchanges and bulk transfer of data.

CoAP web services have been designed for end-to-end constrained devices. A detailed description of CoAP's features is presented in [5], most of which have been implemented in the SpeckSim simulator.

B. Routing Protocols

For the purpose of this study, two classes of protocols have been considered: i) proactive and ii) reactive protocols.

1) Proactive Protocols

Proactive protocols involve the generation and maintenance of routing tables by the nodes in the network. Two protocols that fall under this class are RPL and CTP. Even though CTP nodes do not explicitly maintain routing tables but only a single route towards the root node, it can be classified as a proactive protocol.

a) RPL

RPL has been proposed by the IETF ROLL Working Group as a standard routing protocol for IPv6 routing in WSNs, since existing routing protocols do not satisfy all the requirements for low-power and lossy networks.

RPL organises the network as directed acyclic graphs, starting from the root nodes. It forms a non-transitive, non-broadcast, multiple-access, flexible topology, as described in the IETF draft [1].

b) CTP

CTP is a tree-based collection protocol. When the topology is formed, some of the nodes advertise themselves as root nodes, and the rest of the nodes form routing trees to these roots. CTP is address-free, i.e., a node implicitly chooses a root by choosing a next hop.

2) Reactive Protocols

Reactive protocols do not generate routing tables; instead they build and maintain cache tables based on routing information acquired after route discovery events. Two such protocols are DSR and AODV.

a) DSR

DSR was designed for use in multi-hop wireless networks of mobile nodes. It allows the network to be completely self-organised and self-configuring, without the need for any existing network infrastructure or administration.

The protocol is based on a route discovery and a route maintenance mechanism which operate on demand. It provides loop-free routing, does not send periodic packets of

 ISBN: 978-1-61208-203-5

any kind and supports unidirectional links and asymmetric routes.

b) AODV

AODV is a routing protocol for mobile ad-hoc networks. It uses destination sequence numbers to ensure loop freedom at all times, avoiding problems associated with classical distance vector protocols (such as "counting to infinity").

IV. IMPLEMENTATION DETAILS

Table I summarises the main features of the routing protocols.

TABLE I. ROUTING PROTOCOLS SUMMARY

Characteristics	AODV	DSR	CTP	RPL
Class	Reactive	Reactive	Proactive	Proactive
Tree based	No	No	Yes	Yes
Periodic control messages	Yes (maintains neighbour tables)	No	Yes	Yes
Types of traffic	P2P, P2MP	P2P, P2MP	MP2P, P2MP	P2P, MP2P, P2MP
Types of tables	Routing Table	Cache Table	None (just next hop)	Routing Table
Fault tolerance	Yes	Yes	Yes	No
Communication links	Bidirectional	Unidirectional Bidirectional	Unidirectional	Unidirectional Bidirectional

A. SpeckSim Simulation Framework

The SpeckSim simulation framework [13] is a behavioural level simulator designed for modelling and performance analysis of WSNs. SpeckSim enables modelling and simulation at the different levels of abstraction: devices, networks, layers of the protocol stack, and the application and deployment environment. The simulator incorporates several protocols at the data link and network layers, radio channel models and hardware models for the analysis of power consumption and resource usage.

B. Fire Evacuation from a "Smart" Building

An example of an application explored in this paper is the monitoring and emergency evacuation of a building in the event of a fire, using a internet-enabled WSN attached to the building fabric. The web service identifies the location of the occupants and dynamically computes the safest path [8] towards one of the exits (should such a path exist) and the direction towards this exit is displayed to the occupants in the form of a strobing LED. CoAP (with UDP) is used for this purpose. A reliable transport protocol is not needed due to CoAP's simple retransmission mechanism.

The implementation of the different routing protocols enables effective computation of the hazard times and the safest path from the fire towards the exits. The simulation experiments investigate the impact of the choice of routing protocol. The aim is to examine which one is better suited for this application and the trade-offs in their performance.

C. Choice of SpeckMAC-D as the Data Link Protocol

Our application features low data access rates and the Media Access Protocol (MAC) protocol should be chosen accordingly. The SpeckSim simulator provides a library of MAC protocols from which SpeckMAC-D was chosen as it had outperformed the other protocols in terms of energy consumption and battery lifetime in a previous published study [16]. Unlike other channel probing protocols, SpeckMAC-D performs better for both unicast and broadcast packets which further justifies its selection for this application.

D. CoAP Implementation in SpeckSim

The CoAP implementation in SpeckSim conforms to the description in Draft-8 [3]. CoAP nodes communicate by passing CoAP messages comprising of a fixed 8-byte header. The messages come in different types: Confirmable, Non-confirmable, Acknowledgement, and Reset. Confirmable messages guarantee delivery through the network. They are transmitted in the form of simple retransmissions by increasing the timeout by an order of 2 until the number of maximum retransmissions allowed is reached. For the purpose of the fire evacuation application, all messages are declared to be "Confirmable".

The implementation of the emergency monitoring and evacuation application involves the transmission of CoAP messages at regular intervals between the nodes that detect a person's presence and the exits of the building. The locations of the people within the building are monitored as they traverse the safe paths towards the exits. The paths are continuously updated, taking into account the fire's progress.

E. Implementation of Routing Protocols in SpeckSim

This section briefly describes the implementation of the different routing protocols under study, which are evaluated using CoAP for the fire evacuation application.

1) RPL

The RPL implementation provided in SpeckSim is based on the IETF draft [2]. RPL is optimised for collection networks (ones based on typical traffic of multipoint-to-point (MP2P) and point-to-multipoint (P2MP)), with occasional point-to-point (P2P) traffic.

RPL uses MP2P traffic for data collection and P2MP traffic for configuration purposes. The collection networks have multiple nodes that report periodically to a few collection/sink nodes. Sink nodes rarely choose P2P communication with one of the sender nodes.

2) CTP

CTP is a tree-based collection protocol. When a root node starts up, it broadcasts beacons (routing frames) to generate bidirectional links between the nodes. When a non-root node starts up, it sends routing frames with the "P" bit set (i.e., requesting routing information) until it receives a reply (containing its next hop (parent), the node id, and a metric ETX for evaluating the best parent node). After receiving the reply, it starts broadcasting beacons (similar to the root node) and it can establish connections with new adjacent nodes. Each node holds a parent list as a backup in

 ISBN: 978-1-61208-203-5

case the parent node fails. Should this happen, selection of a new parent node will occur.

The protocol checks for frame duplications and allows up to 32 retransmissions in case of lost data frames or acknowledgements.

The ETX metric was changed to "hop count" due to the ambiguity in the protocol's specification on how to deal with routing loops, thus resulting in loop-free routing.

3) AODV

The implementation of the AODV protocol in the SpeckSim simulator is based on the IETF draft [21].

A node broadcasts a Route Request (RREQ) message when it needs to find a route for a new destination. A route can be obtained when the RREQ either reaches the destination itself or an intermediate node with a 'fresh enough' route to the destination (a 'fresh enough' route is a valid route entry for the destination whose associated sequence number is at least as great as that contained in the RREQ). Each node receiving the request caches a route back to the originator of the request, so that the reply can be unicast from the destination to that originator, or likewise from any intermediate node able to satisfy the request.

Route Error messages are used to propagate link/node failures and changes through the network. The messages may be either broadcasts or unicasts.

4) DSR

The DSR implementation provided in the SpeckSim simulator is based on the paper authored by David B. Johnson et al. [9].

Before the transmission of data packets containing CoAP messages, a node first searches for a route in its cache table. If it finds a route towards the desired destination, it builds the data packet by adding the necessary header information which includes a list/path of nodes that the packet will have to follow in order to reach the destination.

The other nodes in the network will forward the data packet based on the routing information transported in the header of the packet. When a node receives a data packet, it will send an acknowledgement (ACK) message to the node that previously sent the message (one-hop ACKs).

If the node does not find a route towards the desired destination in its cache table, it will initiate the Route Discovery process. The current DSR implementation uses two types of Route Requests: a simple Route Request and a piggyback Route Request (contains the routing information of a Route Reply). This allows it to support unidirectional links and avoid infinite recursion of Route Discoveries.

If the packet is retransmitted for the maximum number of times (15), this will generate Route Error messages which are used to identify the link over which the packet could not be forwarded. The cache table stores only one route for a destination and is populated by the receipt of piggyback Route Requests and Route Replies.

V. RESULTS AND ANALYSIS

We have simulated two building topologies in the SpeckSim simulator: a grid (Manhattan) topology and a less regular topology of a floor in a real building, the Informatics Forum (Figure 1). In both cases each node is within radio range of its immediate neighbours. In Figure 1, the 24-node WSN populates the corridor. Note that the most favourable placement for RPL and CTP root nodes is at the exit nodes of the building, such that the MP2P capability of these protocols can be exploited.

Figure 1. Informatics Forum floor plan and its representation in SpeckSim

A. Test Cases

The fire evacuation application involves simulating people within the building and their passage to safe exits. It also simulates the continuous transmission of CoAP messages between the nodes that detect people and the exit nodes. These messages represent updates on people's location in the building as they move along the paths towards the exits. We have implemented the fire evacuation application on two networks: a 16-node grid-based topology and a 24-node building topology.

The fire evacuation scenario was simulated for the entire protocol stack for the following metrics: delivery ratio, latency, overhead, power consumption, and fault tolerance. Also, scalability studies were performed on grid networks for the following metrics: overhead, packet loss and latency.

B. Results

Each node simulated in SpeckSim has the following characteristics:

- Battery: capacity - 1mAh, voltage- 3V.
- MCU: active current - 0.005mA, sleep current - 0.001mA, off current - 0mA.
- Radio: Perfect Radio Shell, range - 0.35 units.
- Power up delay: Min=0s, Max=1s.

We now present the results that have been gathered by running the fire monitoring and evacuation application in SpeckSim, for the different routing protocols under CoAP-UDP. The results presented are the average of six runs.

AODV exhibits the highest delivery ratio of 100 percent, guaranteeing the transmission of all the messages in the network to the intended destinations.

The latency (Figure 2) is important for the chosen scenario because emergency evacuation requires rapid responsiveness in the network for the short period of time of the evacuation. RPL outperforms the other protocols, as the exit nodes of the building were configured as root nodes, thus leading to effective route selections. The higher latency in AODV and DSR (an average of 11 seconds) can be attributed to their time-consuming route discovery process. However, the 16 second average discrepancy between the two is due to AODV's use of only bidirectional links.

 ISBN: 978-1-61208-203-5

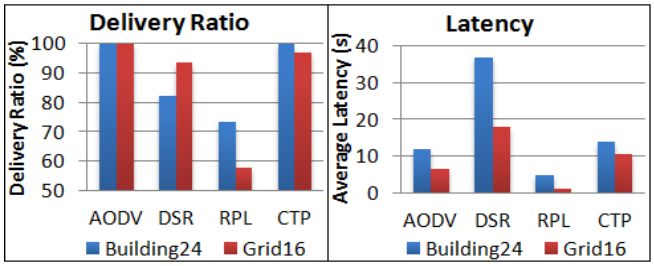


Figure 2. Delivery Ratio and Latency measurements for the two scenarios

The protocol overhead is a useful metric for analysis because it has a direct effect on the average power consumption of the network. It was measured by counting the number of control packets exchanged in the network over a period of time (approximately 650s). We observed a lower overhead for the proactive protocols (Figure 3), owing to the usage of algorithms such as the Trickle timer, which reduce the control packet exchange when the topology is stable.

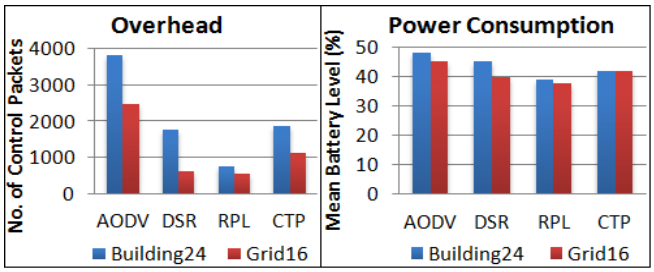


Figure 3. Overhead and Power Consumption results for the two scenarios

AODV needs to maintain neighbour tables in order to use only bidirectional links, but this increases its overhead. Due to the low number of data packets generated (approximately 250 packets) and the short simulation time (approximately 650s), DSR did not have the chance to outclass the other protocols in terms of overhead, even though it does not exchange periodic messages. A more significant difference is shown in Figure 4 of the scalability scenarios, where DSR clearly has lower overhead compared to the other protocols.

Figure 3 also shows the power consumption in terms of the percentage of depleted battery life at the end of the simulation run. This metric is important, as the batteries must last until the evacuation of the building is complete. It was observed that, for all the routing protocols, less than 50% of the batteries' power levels (1 mAh capacity) were drained after running the scenario. Note that any type of battery likely to be used in a real-life deployment is expected to be in order of hundreds of mAh (i.e., CR2032 provides 220 mAh or AA batteries which provide 2500 mAh).

The mean battery consumption was measured when each protocol was run in the simulator for the same type and number of specks. All the protocols displayed similar battery lifetimes/consumption because of the limited run time of the scenario. It may be possible to observe more prominent differences in power consumption if they were simulated on larger topologies for a longer duration.

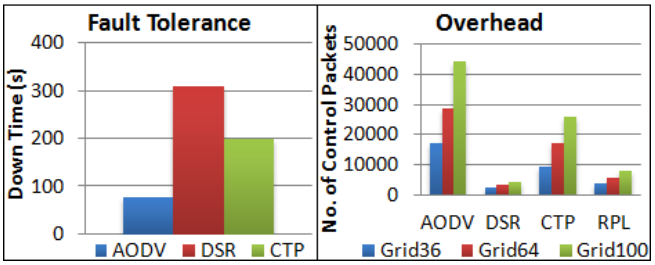


Figure 4. Fault Tolerance (building scenario), Ovehead (grid topologies)

Fault tolerance is another important metric in the context of this application due to the increased chance of the nodes getting damaged. Node failures must be detected and propagated throughout the network so that alternative paths can be found in a timely manner. No fault tolerance mechanisms have been defined in recent RPL drafts because of the overhead that may be created in terms of bandwidth and energy consumption. Therefore, RPL's response to node failures cannot be evaluated.

Figure 4 shows that AODV has the highest tolerance for node failures. This plays an important role in its selection for the CoAP-UDP stack for this particular application.

It can be seen that DSR's down time is higher in comparison to AODV. One plausible explanation is that the DSR implementation in SpeckSim is built to work over both unidirectional and bidirectional links. This implies that the Route Discovery process for this reactive protocol may cause the Route Reply to reach the sender through a different path from that of the Route Request, which causes an additional delay. AODV makes use only of bidirectional links (Route Replies use the backward route of the Route Request), thereby having a reduced down-time. Also, DSR retransmits a packet 15 times before considering a route to be broken, as opposed to AODV which performs only 5 retransmissions.

The protocol drafts do not specify most of the delays and timers used by the protocols, thus making these values implementation specific. Since the intervals for the periodic control packets are implementation specific, in the case of AODV, DSR and CTP when choosing these values, the focus was on reducing the overhead rather than minimising the reconvergence time of the network. This explains the significant down-time of the network when a node fails.

C. *Scalability Results*

The scalability tests have the purpose of validating the results obtained in the case of the fire scenario for the grid and building topologies.

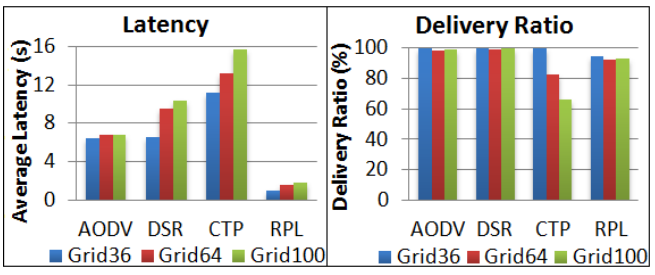


Figure 5. Latency and Delivery Ratio (grid topologies)

 ISBN: 978-1-61208-203-5

The graphs presented in Figures 4 and 5 show that the pattern of the results obtained for the evacuation scenario is maintained as the grid network is scaled to larger topologies.

The only pattern that is not maintained is in the case of the overhead metric for DSR. As previously argued, the protocol is designed not to exchange periodic messages, so it is expected to have a lower overhead over a longer run-time period. For the emergency evacuation scenario DSR did not have the opportunity to outclass the other protocols in terms of overhead. However, a more significant difference is noticed in the scalability studies (Figure 4) where it has a clearly lower overhead than the other protocols.

TABLE II. RESULTS SUMMARY

Characteristics	AODV	DSR	CTP	RPL
Delivery Ratio (%)	100	87.94	96.98	65.62
Average Latency (s)	9.30	27.46	12.21	3.08
Overhead (no. of control packets)	3163	1199	1508	660
Power Consumption (% battery left)	53.32	57.26	58.09	61.50
Fault Tolerance (s) -down time-	78	310	199	-

VI. CONCLUSIONS

This paper has demonstrated an approach for analysing the choice of routing algorithms for a CoAP-UDP protocol stack for internet-enabled WSNs. Table II summarises the performance results for the four routing protocols for the fire evacuation scenario. RPL outperforms the other routing protocols for three out of five metrics. However, it is not fault tolerant and has the lowest delivery ratio.

We can observe that no one protocol outperforms the others for all the metrics which were selected to be relevant to the application. Therefore, the selection of the appropriate protocol to be used with the CoAP-UDP network stack would depend on the weightage accorded to each metric.

In the given scenario, the overhead metric was included to gauge its impact on power consumption. We concluded that power is less of an issue for the time duration simulated. Therefore the overhead metric should not be the prime reason for selecting a routing protocol for this scenario.

Of the five metrics chosen for evaluation one can prioritise three of them: delivery ratio, latency and fault tolerance. One can observe in Table II that AODV and RPL are the two most competitive protocols. Whereas AODV responds well to failures and exhibits a high delivery ratio, RPL has a significantly lower latency.

In case of the fire emergency scenario, the probability of the sensors getting damaged is high. Thus, the network must be able to react to topology changes caused by node failures. Since the current RPL implementation is not fault tolerant, this leaves us to conclude that the most suitable routing protocol (from the ones evaluated) for use in a emergency evacuation scenario is the AODV routing protocol.

ACKNOWLEDGEMENT

The authors wish to thank partial support from ICT FP7 projects HOBNET (257466) and PLANET (257649).

REFERENCES

[1] N. Kushalnagar, G. Montenegro, and C. Schumacher, “IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs): Overview, Assumptions, Problem Statement, Goals”, rfc 4919, 2007.

[2] T. Winter, P. Thubert, and the ROLL Team, “RPL: IPv6 Routing Protocol for Low power and Lossy Networks draft-ietf-roll-rpl-10”, June 28, 2010.

[3] Z. Shelby, K. Hartke, and B. Frank, “Constrained Application Protocol (CoAP) draft-ietf-core-coap-08”, Nov. 1, 2011.

[4] W. Colitti, K. Steenhaut, N. De Caro, B. Buta, and V. Dobrota, “Evaluation of Constrained Application Protocol for Wireless Sensor Networks”, In Proc. of Local & Metropolitan Area Networks (LANMAN), 2011 18th IEEE Workshop on, Oct. 2011.

[5] K. Kuladinithi, O. Bergmann, T. Pötsch, M. Becker, and C. Görg, “Implementation of CoAP and its Application in Transport Logistics”, in Proc. IP+SN, Chicago, IL, USA, 2011.

[6] N. Tsiftes, J. Eriksson, and A. Dunkels, “Low-power wireless IPv6 routing with ContikiRPL”, in Proc. 9th ACM/IEEE International Conference on Information Processing in Sensor Networks, April 12-16, 2010, Stockholm, Sweden.

[7] J. Tripathi, J. C. de Oliveira, and J. P.Vasseur, “A performance evaluation study of RPL: Routing Protocol for Low power and Lossy Networks”, Information Sciences and Systems (CISS), 2010 44th Annual Conference on, March 2010, pp. 1–6.

[8] M. Barnes, H. Leather, and D.K. Arvind “Emergency Evacuation using Wireless Sensor Networks”, in Proc. SenseApp 2007: 2nd IEEE Int. Workshop on Practical Issues in Building Sensor Network Applications, 15 - 18 Oct. 2007, Dublin, Ireland, IEEE.

[9] D.B. Johnson, D.A. Maltz, and J. Broch. “DSR: The Dynamic Source Routing Protocol for Multi-Hop Wireless Ad Hoc Networks”, Ad Hoc Networking, chapter 5: pp. 139 - 172. Addison-Wesley, 2001.

[10] E. Egea-Lopez, J. Vales-Alonso, A. Martinez-Sala, P. Pavon-Marino, and J. Garcia-Haro, “Simulation tools for wireless sensor networks", in Proc. Int. Symp. on Performance Evaluation of Computer and Telecommunication Systems, 2005, pp. 559-566.

[11] F. Osterlind, A. Dunkels, J. Eriksson, N. Finne, and T. Voigt, “Cross-level sensor network simulation with COOJA," in Proc. 31st IEEE Conf. on Local Computer Networks, 2006, pp. 641-648.

[12] P. Levis, N. Lee, M. Welsh, and D. Culler, “TOSSIM: Accurate and scalable simulation of entire TinyOS applications,", 1st Int. Conf. On Embedded Networked Sensor Systems. ACM, 2003, pp. 126 – 137.

[13] Specksim, http://www.specknet.org/dev/specksim, [April 2012].

[14] T. Yang, M. Ikeda, G. DeMarco, and L. Barolli, “Performance Behavior of AODV, DSR and DSDV Protocols for Different Radio Models in Ad-Hoc Sensor Networks”, in Proc. Int. Conf on Parallel Processing Workshops, Sept. 2007.

[15] I.E. Radoi, “Evaluation of Routing Protocols in WSN”, MSc Dissertation, School of Informatics, University of Edinburgh, 2011.

[16] K.J. Wong and D.K. Arvind, “SpeckMAC: low-power decentralised MAC protocols for low data rate transmissions in specknets.” Proceedings of the 2nd international workshop on Multihop ad hoc networks from theory to reality. ACM, 2006. 71-78.

[17] IPSO Alliance, http://www.ipso-alliance.org/, [April 2012].

[18] https://datatracker.ietf.org/wg/core/charter/, [April 2012].

[19] M. Kovatsch, S. Duquennoy, and A. Dunkels. “A Low-Power CoAP for Contiki.”, in Proc of the Workshop on Internet of Things Technology and Architectures (IEEE IoTech 2011), Spain, Oct 2011.

[20] http://sourceforge.net/projects/libcoap/, [April 2012].

[21] C. Perkins, E. Belding-Royer, and S. Das, “Ad hoc On-Demand Distance Vector (AODV) Routing”, rfc 3561, July 2003.

 ISBN: 978-1-61208-203-5

Performance Evaluation of Data Delivery Procedure in IEEE 802.15.4 Based on Discrete-Time Markov-Chain

Peng Hao, Weiting Liu, Jianhua Wang
School of Electronics and Information
Jiangsu Uinversity of Science and Technology
Zhenjiang, 212003, China
Email: peter.haopeng@gmail.com, {lwt_just, jhwang_just}@163.com

***Abstract*—Data delivery procedure (DDP) based on IEEE 802.15.4 involves a series of sub-procedures. They are CSMA/CA (Carrier Sense Multiple Access/Collision Avoidance), data transmission (Tx), acknowledgment (ACK) related behavior (ACK wait duration and ACK transmission). Any failure during this procedure leads to an unsuccessful delivery. This procedure, in fact, determines the network performance, yet not received adequate concern. The algorithm of CSMA/CA, which generally has also been simplified in previous literature. We investigate a discrete-time Markov-chain (DTMC) for DDP without simplification. Due to these sub-procedures, four cases during this procedure are proposed via DTMC models. Particularly, we evaluate the impact of different times of retransmission (ReTx) on the network performance. The performance is investigated in terms of throughput, data delivery ratio and time delay. We also verify our analysis via simulation. Both theoretical and simulation imply that less ReTx can bring better performance.**

Keywords-802.15.4 MAC; CSMA/CA; Data delivery procedure; Discrete-time Markov-Chain; Performance evaluation.

I. INTRODUCTION

Since IEEE 802.15.4 [1] was firstly introduced ten years ago, it has distinguished itself for low data-rate, low cost and low energy consumption. Both academia and industry have devoted great effort to this field.

We note that recent literature has addressed more on some specific IEEE 802.15.4 protocol improvement and applications than the comprehensive performance study itself. Meanwhile, when 802.15.4 MAC performance is concerned, much attention has been focused on CSMA/CA algorithm only, which generally has also been simplified. However, this algorithm is just the beginning of DDP, followed by data Tx, ACK wait duration, and ACK Tx. In this paper, we illustrate a convincing analysis via comprehensive DDP, with unsimplified CSMA/CA. Our work is to evaluate the MAC performance during the procedure of delivering packets between two nodes via one hop.

The rest of the paper is organized as follows. In Section II, we illustrate data delivery procedure and CSMA/CA algorithm in 802.15.4 MAC. In Section III, we overview the related work on performance evaluation of 802.15.4 MAC. In Section IV, discrete-time Markov chain models are proposed for CSMA/CA and DDP. In Section V, the performance is evaluated via both analytical and simulation work. And finally, we summarize our work in Section VI.

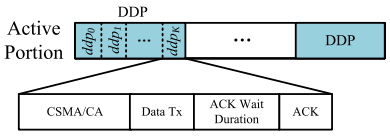


Figure 1. Data delivery procedure in active portion

II. DATA DELIVERY PROCEDURE

IEEE 802.15.4 MAC sublayer provides beacon-enabled and non-beacon-enabled operations. Our attention in this paper is drawn to the beacon-enabled one. Also, the MAC allows the superframe with both active portion and inactive portion [1]. In inactive portion, the node turns into sleep mode and no data is delivered. we assume that only active portion is available since the maximum performance is concerned in this paper. While in active portion, data packets are delivered via DDP. As shown in Figure 1, each DDP involves $(macMaxFrameRetries + 1)$ times of sub-DDP, namely, ddp. The parameter, $macMaxFrameRetries$, implies the maximal number of retransmission of the packet [1]. For simplicity, we use K and k to indicate $macMaxFrameRetries$ and the times of ddp, respectively, namely, ddp_k, where $k = 0, \cdots, K$. Data packets shall be delivered if ddp_k is successfully carried out, involving CSMA/CA, Data Tx and ACK. In addition, there is a constant, IFS (Inter-Frame Space) [1], between the successful delivered data and the consecutive delivery. It can be neglected and not considered in this paper. Any failure of ddp_k results in ddp_{k+1}. All probabilities in this paper are assumed to be obtained based on the steady state.

The parameters in DDP are set by MLME (MAC Layer Management Entity). After the data delivery is notified by MLME, the times of ReTx, k, is initialized to be zero, data shall be maximally repeated $(K + 1)$ times of ddp, until `SUCCESS` is made. Otherwize, `FAILURE` is notified. In other words, K is one of the key factors to determine the performance.

 ISBN: 978-1-61208-203-5

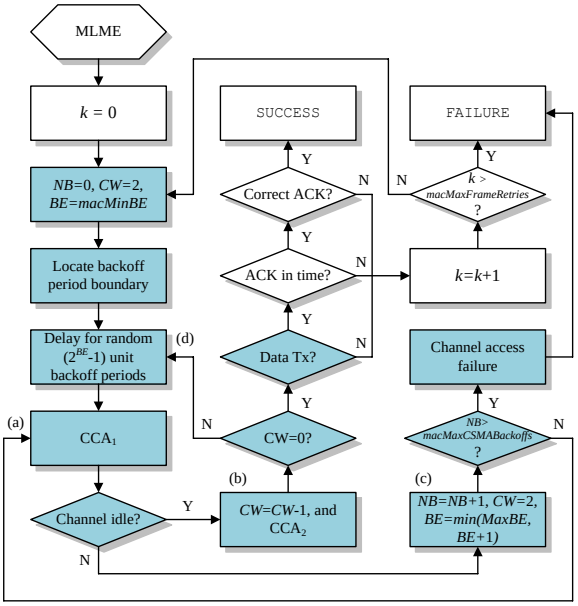


Figure 2. Data delivery procedure, including CSMA/CA algorithm (in shade boxes)

The mechanism of CSMA/CA is the key component in 802.15.4 MAC, shown in the shade boxes in Figure 2. It is adopted to arrange the nodes in the network with an appropriate order when they access the channel. It starts from the notification issued by MLME, and ends when the channel is found either idle or busy. In brief, two behaviors are involved in this algorithm, backoff period (BP) delay and twice CCAs (Clear Channel Assessment). Herein, for the simplicity, we denote them as once $BP-CCA-CCA$. The following gives the details.

After the k initialization, the node firstly perform a ***BP***, as shown in Figure 2. The length of ***BP*** is a random value based on the period determined by BE, namely, $(2^{BE}-1)$ units of $aUnitBackoffPeriod$, that is, 20 symbols. Then MAC starts to count down the time prior to the first CCA.

Then the first CCA shall be performed at the boundary of the backoff period, as (a) implies in Figure 2. If the channel is accessed idle, CW self-deceases by one and CCA shall be performed again (see (b) in Figure 2). If this second CCA successfully finds the channel idle, then the CSMA/CA is successful and data shall be transmitted.

However, if the channel is found busy at the first CCA (see (c) in Figure 2), MLME enables the next ***BP*** with a new length, determined by an updated BE, where $BE = min(MaxBE, BE+1)$. Or if the channel is idle in CCA_1, while busy in CCA_2, namely $CW = 0$, then MLME activates a new first CCA in the next round of ***BP*** (see (d) in Figure 2). If both the twice CCAs find the channel busy, a notification of `FAILURE` is indicated by MLME and forwarded to the *Upper Layers* [1].

CSMA/CA algorithm consists of NB times of $BP-CCA-CCA$ procedures, where the length of BP is determined by BE, and the potential number of CCA is determined by CW, as follows,

- NB: the number of times the CSMA/CA algorithm shall be required to backoff while attempting the current transmission. $0 \leq NB \leq macMaxCSMABackoffs$, where $1 \leq macMaxCSMABackoffs \leq 5$, but the default value is 4. NB is initialized 0. In our work, we use Q and q to denote $macMaxCSMABackoffs$ and the number of times, namely, $q = 1, \cdots, Q$, where $Q = macMaxCSMABackoffs$.
- CW: the contention window length, defining the number of backoff periods that need to be cleared of channel activity before the transmission can commence. In slotted CSMA/CA, by default, the length is set to be 2, namely twice CCA.
- BE: the backoff exponent. It is related to the length of backoff period a node shall wait before attempting to access a channel. The value depends on battery life extension or not, as shown in Figure 2. Here we assume $BE = macMinBE$, where $0 < macMinBE \leq 3$.

III. Related Work

The community has been evaluating the performance of 802.15.4 MAC by simplifying CSMA/CA algorithm (for example, only once CCA in [2], [3]). BE, CW and NB have mostly received specific attention, so has the payload size of data frame, N_{MSDU}. The impact of BE and N_{MSDU} are concerned in [4]–[9], where different methods have been proposed to determine the length of ***BP***. CW is investigated in [2], [8], which concludes a large number of CCA can lead to less throughput. Ramachandran et al. [2] also evaluates the influence of NB. By focusing on CSMA/CA, these methods above claim to involve the whole data transmission procedure. However, this might no be true since CSMA/CA is the beginning of the procedure. ACK and retransmission (ReTx) also need to be concerned.

Quite limited literature has considered the impact of ACK during the data transmission. Much work shows their interest in the difference between with and without ACK. Mišić starts one of the most pioneering work in ACK-related 802.15.4 MAC performance evaluation. The fruitful research has been accomplished in this field including different topology network with ACK (star [10] and cluster [11]), and different transmission in terms of uplink/downlink [12]. However, as mentioned in [2], [13], the analytical models diverse from their simulation results. Reference [13], [14] also concerns the up-link transmission respectively with/without ACK. Particularly in [13], an accurate and scalable analytical model is proposed. However, their work may not be comprehensive enough since only successful ACK is involved.

 ISBN: 978-1-61208-203-5

The work on the impact of retransmission on the network performance is still inadequate. The proof of applying DTMC in the evaluation work has been presented in [3]. Three dimensional DTMC is proposed in [15], [16], where the number of retransmission is taken into account. However, the data delivery procedure in their work might not be illustrated appropriately. Finding channel busy at the first CCA leads to the current CSMA/CA, again. This, in fact, should result in the next CSMA/CA procedure if the maximum times (namely K) of retries have not been met. Jung also proposes a three-type DTMC model to analyze the performance [17]. By considering the inactive portion in the superframe, their work in fact focuses on the unsaturated network. Their contribution of DTMC also includes the probability of deferring the data frame that can not be completed in current superframe to the next superframe. However, in a saturated situation (which is concerned in our work), their method might not be applicable. Though having the impact of different number of retransmissions considered, as mentioned in [13], Jung's work may increase the complexity of the analysis and limits the scalability.

There are also other factors that can affect the performance, including the number of nodes involved in the network, signal fading and interference, and so on. Our previous work has investigated the impact of the number of nodes with both star [18] and tree topology [19], [20], respectively. Channel interference is concerned in this paper, and signal fading will be evaluated in our future work.

In our work, we propose a comprehensive DTMC for 802.15.4 MAC. Our attention is focused on the impact of the maximum number of retransmission, K (namely, $macMaxFrameRetries$). The performance is investigated in terms of network throughput, packet delivery ratio and time delay.

IV. DTMC OF DDP

We illustrate the whole procedure of DDP in Figure 2 via stochastic analysis in terms of the procedure of ddp_k, as shown in Figure 3. The procedure is initialized by $k = 0$ (namely ddp_0). and K times of ReTx (namely ddp_k, $k = 1, \cdots, K$). Each of them involves at most Q times (namely $amacMaxCSMABackoffs$) of $BP - CCA - CCA$, followed by once Tx and once ACK. As shown in this figure, the subscript k in BP, CCA, Tx and ACK indicates the k-th ddp; the subscript, q in BP and CCA, denotes the q-th $BP - CCA - CCA$ procedure; and respectively, i and $i|i$ depict channel $idle$ (in the first CCA) and channel $idle$ at the second CCA, given $idle$ in the first CCA. Also, the superscripts, n and c depict node and channel, respectively.

- ddp_k: the k-th procedure of data delivery. This procedure includes Q times of CSMA/CA, once data transmitting (namely Tx_k) and the behavior of waiting for and processing ACK (namely ACK_k). There are algotether $(K + 1)$ times of ddp_k.

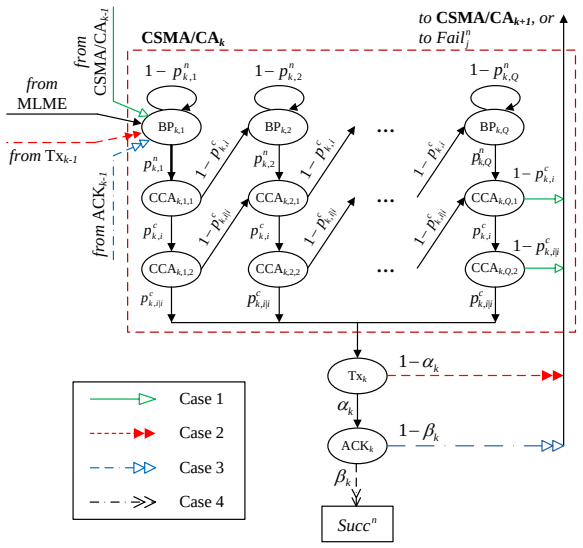


Figure 3. DTMC of DDP and four cases in DDP

- $BP_{k,q}$: the q-th $Backoff\ Period$ in the ddp_k.
- $p^n_{k,q}$: the probability for the node to perform the $BP_{k,q}$.
- $CCA_{k,q,v}$: the v-th CCA in q-th $BP - CCA - CCA$ procedure in ddp_k, where $1 \le v \le CW$, $1 \le q \le Q$. In our work, CW is initialized to be 2.
- $p^n_{k,i}$: the probability that the channel is found $idle$ at the first CCA (namely $CCA_{k,q,1}$), here i denotes $idle$.
- $p^c_{k,i|i}$: the probability that the channel is found $idle$ at the second CCA (namely $CCA_{k,q,2}$), given the $idle$ channel in $CCA_{k,q,1}$.
- α_k: the probability that the transmitting in the PHY sublayer is successful, considering the channel noise or interference.
- β_k: the probability that the correct ACK is received in time.

A. Four Cases in Data Delivery Procedure

We understand CSMA/CA, data Tx in PHY sublayer and ACK-related process can all impact the network performance. Therefore, we take all of them into account, as shown in Figure 3. In each ddp, if the current CSMA/CA is unable to lead to Data Tx, another CSMA/CA in a new ddp shall be processed, which can also be activated by the failure of Data Tx in PHY sublayer. Additionally, if the ACK-related process fails, Data Tx in the new ddp shall be carried out again in the PHY sublayer.

The data is successfully delivered if and only if the **Correct** ACK is received within $macAckWaitDuration$, namely 54 symbols [1]. A notification of `SUCCESS` is generated by MLME. Otherwise, a notification of `FAILURE` occurs. These four cases are,

- Case 1: unsuccessful data transition due to the failure of CSMA/CA; or

- Case 2: unsuccessful data transition, due to the channel failure (noise or interference); or
- Case 3: the data is successfully transmitted, but **No** ACK is received within the certain period of time ($macAckWaitDuration$ symbols); or, received in time, but the ACK is **Incorrect**. In other words, the DSN (Data Sequence Number) this ACK contains is not same with the one from the data or MAC command that is being acknowledged [1]; or
- Case 4: successful data transmission and correct ACK received in time.

Based on Figure 3, we can have the probabilities for Case 1 - Case 4 in DDP, denoted as $p^n_{c_1}, \cdots, p^n_{c_4}$, respectively, as follows,

$$p^n_{c_1} = \sum_{k=0}^{K}[(1-p^c_{k,i})\cdot\pi(CCA_{k,Q,1}) + (1-p^c_{k,i|i})\cdot\pi(CCA_{k,Q,2})], \quad (1)$$

$$p^n_{c_2} = \sum_{k=0}^{K}(1-\alpha_k)\cdot\pi(Tx_k), \quad (2)$$

$$p^n_{c_3} = \sum_{k=0}^{K}(1-\beta_k)\cdot\pi(ACK_k), \quad (3)$$

$$p^n_{c_4} = \sum_{k=0}^{K}\beta_k\cdot\pi(ACK_k). \quad (4)$$

In addition, we assume the data to be transmitted at each node is subject to Poisson process, with the mean as p. Also, p is the normalized traffic load prior to DDP. And the state of $Fail^n_j$ denotes the j-th failure in DDP, where $j = 1, 2$.

Also the parameter of $p^n_{k,q}$ is assumed as a geometric random variable [2], [3], as shown in (5) [1]. This is consistent with the fact that the lower value of the q can lead to the bigger chance to perform *BP-CCA-CCA*. Furthermore, the BP can be regarded *memoryless*. Meanwhile, the probability of channel failure due to interference or noise, namely $1-\alpha_k$, is also assumed to be subject to the uniformly distributed white noise with $0.8 \le \alpha_k \le 1$, where $k = 0, \cdots, K$.

$$p^n_{k,q} = \frac{1}{\frac{2^{BE}-1}{2}+1} = \begin{cases} \frac{2}{2^{q+2}+1}, & \text{if } q = 1, 2; \\ \frac{2}{2^Q+1}, & q = 3, \cdots, Q. \end{cases} \quad (5)$$

Meanwhile, we note that CCA behavior is actually independent to the procedures of data delivery because CCA is determined by the channel state. In other words, The probability of $p^c_{k,i}$ is assumed to be the same at different ddp_k. Therefore it is rewritten as p^c_i. And so is the probability of $p^c_{k,i|i}$, rewritten as $p^c_{i|i}$. Moreover, since all the probabilities in our DTMC is assumed to be obtained in the steady state, we use π to denote the steady state, followed by the MAC behavior or channel state. Therefore, the steady state of BP, *CCA*, Tx, ACK, $Fail$, and $MLME$ can be obtained.

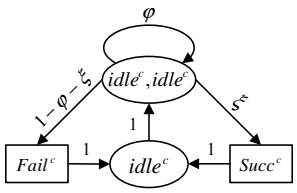


Figure 4. Channel DTMC

B. DTMC of Channel

The physical channel plays a vital role in the evaluation. This is not only because of the potential noise and interference which has been considered in Case 2 in Section IV-A, but also the fact that the channel states (*idle* or *busy*) determine whether the data transmission can be carried out in the first place.

Furthermore, we understand the fact that all frames, including data, command, and ACK, are transmitted via the physical channel. Therefore, the throughput of the network can actually be equal to the throughput of the channel. This can be more reliable than the throughput at nodes or the network coordinator. Unlike the latter which has been adopted in most work in this field, we investigate the network performance via the evaluation on the channel.

Two states are shared by both nodes and the channel, namely $idle$ and $(idle, idle)$. Regarding twice CCAs, $idle^c$ and $(idle^c, idle^c)$ are introduced to facilitate the modeling by combining both MAC process and channel states, where the superscript c refers to $channel$, as shown in Figure 4. The probabilities are illustrated in (6) and (7) respectively.

$$p^c_i = \frac{2-\varphi}{1+(N_{MSDU}+1)(1-\varphi)}, \quad (6)$$

$$p^c_{ii} = \frac{1}{1+(N_{MSDU}+1)(1-\varphi)}. \quad (7)$$

For the channel, these two states lead to the result of either `SUCCESS` or `FAILURE`, denoted by $Succ^c$ and $Fail^c$ in Figure 4. Here, the number of times is not applied to the latter, unlike j in $Fail^n_j$. Assume M source nodes in the network, the probability from $(idle^c, idle^c)$ to $Succ^c$ is,

$$\xi = Mp^n_{t|ii}(1-p^n_{t|ii})^{M-1}, \quad (8)$$

owing to that each time only one source node can successful transmit data frames via the channel. Also, staying at $(idle^c, idle^c)$ means no transmission from source nodes in the network, namely,

$$\varphi = (1-p^n_{t|ii})^M, \quad (9)$$

where $p^n_{t|ii}$ is the probability of transmitting the packet after the successful twice CCA. This parameter is obtained by (10)

$$p^n_{t|ii} = \frac{p^n_t}{p^c_{ii}}. \quad (10)$$

Particularly, we have $0 \leq p^c_{t|ii} \leq 1 - \sqrt[M]{\frac{2}{N_{MSDU}+1}}$, considering both $\varphi > 0$ and $p^n_{t|ii} > 0$.

V. Numerical Results

We assume the network be comprised by thirty homogeneous source nodes, namely $M = 30$. They are sending data to a coordinator node. Based on ns-2, the distance between the nodes and the coordinator is randomly distributed within the working range (15 m) so that the nodes can talk to the coordinator with a single hop. These nodes need to deliver a packet of 100-byte MSDU (namely $N_{MSDU} = 100$ byte) each time with normalized traffic load p. Particularly, since three CSMA/CA-related parameters, BE, CW and NB, have been sufficiently studied by the research community, their impact shall not be addressed. Our work emphasizes on the impact of K ($macMaxFrameRetries$). We evaluate the network performance in terms of network throughput, time delay and packet delivery ratio, as follows:

- $thpt$: the effective network throughput. Namely, the ratio of the MSDU received at the coordinator to the consumed time. In particular, SHR ($Synchronization\ Header$), PHR ($PHY\ Header$), MHR ($MAC\ Header$) and MFR ($MAC\ Footer$) are not concerned, neither the command frames such as beacons [1].
- t_{delay}: the average time consumed when a packet is successfully transmitted from the source node to the coordinator.
- η: the packet delivery ratio. That is, the ratio of the number of MSDU received at the coordinator to the one sent from the nodes.

We begin the evaluation with the probability of Case 1 to Case 3 (refer to (1), (2) and (3)), shown in Figure 5.

In Figure 5, we observe that Case 1 brings the prominent impact to the network. That is, ReTx occurs mostly due to the busy channel during twice CCA. Particularly, the first CCA has a stronger impact on the performance, because the probability of the CCA_1 is much larger than the one of CCA_2, namely, $\frac{\sum_{k=0}^{K}\sum_{q=1}^{Q}\pi(CCA_{k,q,1})}{\sum_{k=0}^{K}\sum_{q=1}^{Q}\pi(CCA_{k,q,2})} \gg 1$.

Obtaining $thpt$ via the channel states can also be found in [2], [3]. However, only CSMA/CA is involved in their work. The throughput in [2], [3] is actually based on the successful transmission at source nodes, rather than channel-based analysis in our work. Also, their throughput involves data packets, ACK frame, beacon frame and other maintenance frames, which did not depict the effective throughput contributed by MSDU. In our work, we are concerned

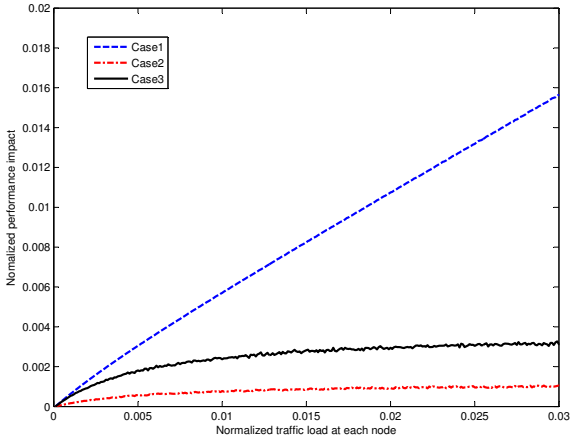


Figure 5. Normalized performance impact of case 1 to case 3

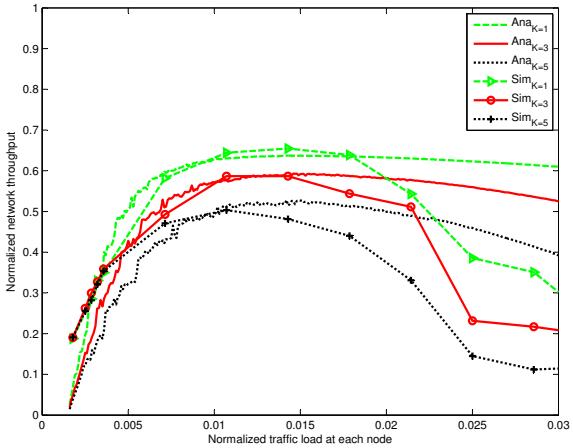


Figure 6. Normalized network throughput at different K

about the throughput due to the data packets only, and recognize that channel noise/interference and ACK-related cases should also be considered. Therefore, given $\pi(Succ^c)$, the probability of Case 4, i.e. $p^n_{c_4|s^c}$, is,

$$p^n_{c_4|s^c} = \frac{p^n_{c_4}}{(K+1)\cdot\pi(Succ^c)}. \quad (11)$$

Now the network throughput is obtained, as illustrated in (12). The throughput at different $maxMaxFrameRetries$ (i.e. K) is shown in Figure 6 .

The analytical results are also verified by the simulation results based on $ns-2$. First, more procedures of ddp_k bring less throughput, because these procedures prolong the packet delivery time. Second, throughput behaves with a saturation interval, as shown in Figure 6. After the saturation, throughput decreases. This is because more packets may have been dropped due to collision.

The total time delay, denoted by t_{sum}, is obtained, as follows,

 ISBN: 978-1-61208-203-5

$$thpt = p^n_{c_4|s^c} \cdot \frac{N_{MSDU} \cdot \pi(Succ^c)}{\pi(idle^c) + \pi(idle^c, idle^c) + N_{MSDU} \cdot \pi(Succ^c) + N_{MSDU} \cdot \pi(Fail^c)} = p^n_{c_4|s^c} \cdot \frac{N_{MSDU} \cdot M \cdot p^n_{t|ii} \cdot (1 - p^n_{t|ii})^{M-1}}{1 + (N_{MSDU} + 1)(1 - (1 - p^n_{t|ii})^M)}. \quad (12)$$

$$t_{delay} = \sum_{k=0}^{K} t_k, \quad (13)$$

where t_k is the time consumed during the ddp_k procedure, as shown in (14),

$$t_k = \sum_{q=1}^{Q} \{\pi(BO_{k,q}) \cdot \tau_{BP} + \sum_{v=1}^{2} [\pi(CCA_{k,q,v}) \cdot \tau_{CCA}]\} + \pi(Tx_k) \cdot \tau_{Tx} + \pi(ACK_k) \cdot \tau_{ACK}, \quad (14)$$

where τ_{CCA}, τ_{Tx}, τ_{ACK} and τ_{BP} are illustrated in (15) to (19), respectively.

$$\tau_{CCA} = 8 \cdot 0.016 = 0.128 \text{ ms}, \quad (15)$$

$$\tau_{Tx} = \frac{MSDU}{Datarate} = \frac{100 \cdot 8}{250} = 3.2 \text{ ms}, \quad (16)$$

$$\tau_{ACK} = macAckWaitDuration + t_{ACK} \quad (17)$$

$$= 0.864 + 0.352 = 1.216 \text{ ms}. \quad (18)$$

$$\tau_{BP} = \begin{cases} 0.32 \cdot (2^{BE} - 1), & if\ macMinBE \le BE \le 4; \\ 0.32 \cdot (2^{aMaxBE} - 1), & if\ BE > 4. \end{cases} \quad (19)$$

where 0.32ms is the length of $aUnitBackoffPeriod$. And t_{ACK} means the time to process the received ACK. By varying K, we have Figure 7.

When K becomes higher, the node spends more time on delivering the packet to the coordinator. Furthermore, the time delay increases significantly at higher traffic load due to collision.

The packet delivery ratio η is illustrated in (20). Also we investigate the evaluation by setting different K, as shown in Figure 8.

$$\eta = \frac{thpt}{M \cdot p \cdot N_{MSDU}} = \frac{p^n_{t|ii} \cdot (1 - p^n_{t|ii})^{M-1}}{1 + (N_{MSDU} + 1)(1 - (1 - p^n_{t|ii})^{M-1})} \quad (20)$$

Similar results are obtained in this figures as well. Both analytical work and simulation share the result that delivery ratio is performed in a decreasing trend along the increment of the traffic load. Our simulation also shows that the network relatively keeps enjoying the high ratio when the traffic load is fairly small. It reminds that the packet delivery ratio suffers more at higher traffic loads.

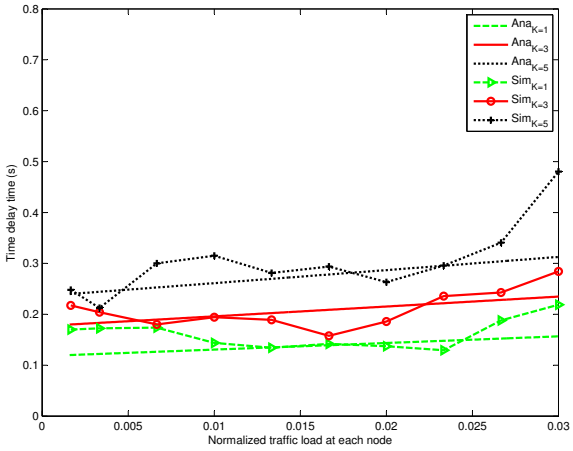


Figure 7. Time delay at different K

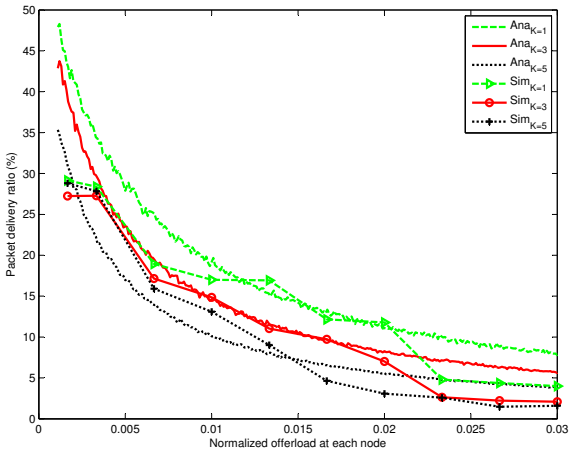


Figure 8. Packet delivery ratio at different K

VI. Conclusion

We proposed discrete-time Markov chain (DTMC) models for the comprehensive data delivery procedure (DDP) in 802.15.4-based beacon-enabled network. DDP includes $(macMaxFrameRetries + 1)$, namely, $(K + 1)$ times of sub-DDP (that is, ddp_k, $k = 0, \cdots, K$). Each ddp_k involves three MAC behaviors. They are standard slotted CSMA/CA algorithm which is comprised of up to

$macMaxCSMABackoffs$ times of backoff periods (BP) and twice CCA, the transmission (Tx) in PHY sublayer and ACK-related (Acknowledgment) process as well. The successful data delivery indicates that, during a DDP, the data packet is transmitted after success in all three behaviors. Because of the success/failure of the three MAC behaviors, four cases are proposed regarding different outcomes of data delivery. Based on the DTMC and the simulation work via ns-2, we evaluate the MAC performance of the network. By varying K, the impact on the network performance are studied, in terms of throughput, time delay and packet delivery ratio global. Our work reveals more K can bring poor performance.

ACKNOWLEDGMENT

The authors would like to thank The University of Melbourne and National ICT Australia Ltd. for the valuable contribution to our work.

REFERENCES

[1] "IEEE standard for information technology - telecommunications and information exchange between systems - local and metropolitan area networks - specific requirements. part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (LR-WPANs)," Tech. Rep., 1 October 2003.

[2] I. Ramachandran, A. K. Das, and S. Roy, "Analysis of the contention access period of ieee 802.15.4 mac," *ACM Transactions on Sensor Networks*, vol. 3, no. 1, pp. 1–29, Mar 2007.

[3] M. Martaló, S. Busanelli, and G. Ferrari, "Markov chain-based performance analysis of multihop ieee 802.15.4 wireless networks," *Performance Evaluation*, vol. 66, pp. 722–741, 2009.

[4] B. Gao and C. H. L. Jiang, "Modeling and analysis of ieee 802.15.4 csma/ca with sleep mode enabled," in *In Proceedings of the 11th IEEE Singapore International Conference on Communication Systems*, Singapore, Nov 2008, pp. 6–11.

[5] J. Gao, J. Hu, and G. Min, "A new analytical model for slotted ieee 802.15.4 medium access control protocol in sensor networks," in *In Proceedings of the Communications and Mobile Computing (CMC'09)*, vol. 2, Yunnan China, Jan 2009, pp. 427–431.

[6] H. Wen, C. Lin, Z.-J. Chen, H. Yin, T. He, and E. Dutkiewicz, "An improved markov model for ieee 802.15.4 slotted csma/ca mechanism," *Journal of Computer Science and Techonology*, vol. 24, no. 3, pp. 495 – 504, May 2009.

[7] X. Ling, Y. Cheng, J. Mark, and X. Shen, "A general analytical model for the ieee 802.15.4 contention access period," in *In Proceedings of IEEE Wireless Communications and Networking Conference*, Hong Kong, March 2007, pp. 316–321.

[8] Z. Tao, S. Panwar, D. Gu, and J. Zhang, "Performance analysis and a proposed improvement for the ieee 802.15.4 contention access period," in *In Proceedings of Wireless Communications and Networking Conference (WCNC'06)*, vol. 4, Las Vegas, USA, Apr 06, pp. 1811–1818.

[9] B. Lauwens, B. Scheers, and A. V. de Capelle, "Performance analysis of unslotted csma/ca in wireless networks," *Telecommunication Systems*, vol. 44, no. 1-2, pp. 109–123, Jun 2010.

[10] J. Mišić, S. Shafi, and V. B. Mišić, "The impact of mac parameters on the performance of 802.15.4 pan," *Ad Hoc Networks*, vol. 3, pp. 509–528, 2005.

[11] J. Mišić, "To acknowledge or not to acknowledge: The case of interconnected 802.15.4 clusters," in *In Proceedings of the 16th Mobile and Wireless Communications Summit*, Budapest, Jul 2007, pp. 1–5.

[12] J. Mišić, S. Shafi, and V. B. Mišić, "Performance of a beacon enabled ieee 802.15.4 cluster with downlink and uplink traffic," *IEEE Transactions on Parallel and Distributed Systems*, vol. 17, no. 4, pp. 361–376, Apr 2006.

[13] J. He, Z. Tang, H.-H. Chen, and Q. Zhang, "An accurate and scalable analytical model for ieee 802.15.4 slotted csma/ca networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 440–448, Jan 2009.

[14] S. Pollin, M. Ergen, S. C. Ergen, B. Bougard, F. Catthoor, A. Bahai, and P. Varaiya, "Performance analysis of slotted carrier sense ieee 802.15.4 acknowledgment uplink transmission," in *In Proceedings of IEEE Wireless Communications and Networking Conference*, Las Vegas, USA, Mar/Apr 2008, pp. 1559–1564.

[15] P. K. Sahoo and J.-P. Sheu, "Modeling ieee 802.15.4 based wireless sensor network with packet retry limits," in *In Proceedings of the 5th ACM symposium on Performance evaluation of wireless ad hoc, sensor, and ubiquitous networks*, New York, USA, Oct 2008, pp. 63–70.

[16] J. Park and S. Sahni, "An online heuristic for maximum lifetime routing in wireless sensor networks," *IEEE Transactions on Computers*, vol. 55, no. 8, pp. 1048–1056, Aug 2006.

[17] C. Y. Jung, H. Y. Hwang, D. K. Sung, and G. U. Hwang, "Enhanced markov chain model and throughput analysis of the slotted csma/ca for ieee 802.15.4 under unsaturated traffic conditions," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 1, pp. 473–478, Jan 2009.

[18] W. Qiu, P. Hao, and R. J. Evans, "An efficient self-healing process forzigbee sensor networks," in *Proceedings of The 7th International Symposium on Communicationsand Information Technologies*, Sydney, Australia, Oct 2007, pp. 1–6.

[19] P. Hao, W. Qiu, and R. J. Evans, "Performance evaluation of ieee 802.15.4 mac in beacon-enabled tree-topology wireless sensor networks," in *in Proceedings of the Fifth International Conference on Systems and Networks Communications*, Nice, France, Aug 2010, pp. 1–6.

[20] W. Qiu, E. Skafidas, and P. Hao, "Enhanced tree routing for wireless sensor networks," in *Ad Hoc Networks, Elsevier*, vol. 7, 2009, pp. 638–650.

 ISBN: 978-1-61208-203-5

Merging Grid into Clustering-based Routing Protocol for Wireless Sensor Networks

Ying-Hong Wang, Yu-Wei Lin, Yu-Yu Lin, Hang-Ming Chang
Computer Science and Information Engineering
Tamkang University
Tamsui Taipei, Taiwan, R.O.C.
inhon@mail.tku.edu.tw, harry040@hotmail.com, jerry198926@hotmail.com, chmcicel7473@gmail.com

Abstract- **Wireless sensors have a finite amount of energy and cannot be recharged after deployment. Therefore, the efficiency of the algorithm is an important consideration. An effective algorithm can reduce energy consumption and prolong network lifetime. In this paper, we proposed a cluster-based routing protocol. First, this algorithm divides networks into several units and those units would be regarded as clusters. Second, according to the remaining energy of the nodes, the nodes will execute cluster-head selection. Finally, routing tables are created for routing within clusters and between clusters. We then compare the method we proposed with others by simulation, and the conclusion proves the method we proposed saves more power resulting in a longer network lifetime.**

Keywords - Cluster; Power saving; Routing protocol; WSNs.

I. INTRODUCTION

In recent years, wireless application and wireless communication markets have become more popular due to the rapid development of wireless communications technology. Moreover, the advances in micro technology has led to wide adoption of multiple wireless network technologies consisting mainly of wireless sensor networks [1], Wireless Local Area Network (WLAN), Worldwide Interoperability for Microwave Access (WiMAX), Ad Hoc Networks, Bluetooth Wireless Personal Area Network(WPAN), etc. In wireless sensor networks, micro-manufacturing technology continues to increase capabilities in environmental sensing, information processing, wireless communications, computing ability, and storage capacity. In order to take full advantage of these advances, reducing energy consumption to extend network lifetime of wireless sensors is critical, and thus an important topic for research.

The characteristics of wireless sensor design calls for small footprint, low cost, power saving, and accurate sensing ability. Not only should the hardware experience a breakthrough in growth, but so should the accompanying software The current areas of research can be divided into the following several categories: routing protocol, target tracking, locating, data aggregation, fault tolerance, sensor node deployment and energy management. Each sensor node has data processing, communication, and data sensing responsibilities--all consuming a limited energy resource. In this premise, a wireless sensor node achieves the greatest benefit from an increase in energy consumption efficiency. Therefore, how to design an effective routing protocol is a very important topic. In our proposal, the major issues is studying reduce energy consumption and routing protocol. In the wireless sensor network applications using the environment as a static target and more under consideration, we hope that the node does not have too strong computing power and other additional equipment in order to achieve as much as possible to reduce energy consumption and cost effectiveness. To achieve this goal we propose a cluster- based routing protocol for wireless sensor networks, that is Merging Grid into Clustering-based Routing Protocol (MGCRP) for Wireless Sensor Networks. Through our proposed method we will show improved energy consumption efficiency resulting in extended network lifetime.

The rest of this paper is organized as follows: Section II presents the related work. Section III elaborates the protocol. Simulation results are discussed in Section IV and we conclude our paper in Section V.

II. RELATED WORK

There are numerous papers on using routing protocols to make wireless sensor networks stable, effective and power saving. Al-Karaki and Kamal [2] and Qiangfeng and Manivannan [3] introduced the concept of using routing protocols in wireless networks.

Recently, there are three leading ways of routing. They are chain-based, cluster-based and tree-based; our paper will focus on cluster-based. In cluster-based routing nodes are divided into clusters and the cluster head will send the data collected from normal nodes to sink. Our research consists of two parts: How to effectively cluster nodes, and how to determine the optimal routing path.

Low Energy Adaptive Clustering Hierarchy (LEACH) is proposed by Heinzelman [4]. This routing protocol divides nodes into several clusters by their location, and the nodes can only communicate with in the same cluster.

A special node will be elected as the cluster head. It will collect data from other normal nodes and then send to the sink. Transmission is the source of large energy consumption, so to ensure equal expenditure of energy by the nodes in the network, another cluster head is chosen after the transmission finishes. However, the cluster-head is chosen at random, so it is hard to determine whether the cluster heads are distributed evenly in the network. Also, in this algorithm, distance between the cluster head and the sink is not considered leading to a potential waste of energy if the distance to the sink is exceedingly far from the cluster head which may be further exacerbated if the randomly chosen cluster head belongs to a high node density cluster.

 ISBN: 978-1-61208-203-5

Energy-Balanced Chain-cluster Routing Protocol (EBCRP) [5] is proposed by Bao Xi-Rong. It is a cluster-based distributed algorithm that builds a path of chains through the uses of a ladder algorithm. EBCRP can be divided into three parts, chain-cluster formation, cluster-head selection and steady state. In chain-cluster formation stage, it divides the network into several rectangular blocks and use a ladder algorithm to build a chained path. The next hop in a ladder algorithm is the next increment along the y-axis.

In the cluster head selection stage, a few nodes will be selected to communicate with the sink. In each round, the node closest to the sink with the most residual energy will be the cluster head. In the steady-state stage, the cluster head will collect data and send it to the sink. Each round the cluster head will change to reduce the load of the cluster head. In this algorithm, every node besides the cluster-head transmits data to their neighboring node resulting in a duplication of data communication between 90% of nodes.

III. MERGING GRID INTO CLUSTERING-BASED ROUTING PROTOCOL

Our proposed routing protocol is divided into two phases: Clustering Phase and Routing Phase. We will add Cluster-Head Rotation Mechanism to maintain routing persistence in the Routing Phase. The Clustering Phase starts after the deployment of sensor nodes. In this phase, the sink through the use of the location mechanism, determines the location of each sensor node by using user defined N value of grid length to divide the network into several grids of the same size and, then calculates the each center of grid and the number of nodes within each grid. Moreover, each grid as a cluster, then merge these valid clusters. After clustering the network, a cluster head is chosen for each cluster, and sends it to all cluster heads in network so that each cluster will have information of other cluster head. After all the cluster heads have been selected, the next step is the transmission stage. But before data transfer can proceed, efficient routes needs to be determined. In this routing phase, we initially only used Bellman-Ford shortest path algorithm by D. Bertsekas and R. Gallager [6] to build the inner-clusters and outer-clusters initial routes. However, this operation consumed too much energy during the sensing and transmit stages. To alleviate this issue, during the data transmission stages we added the cluster head rotation mechanism, to avoid overloading any single node.

A. Network Environment and Assumption

We assume the wireless sensor network is composed of a sink and a large number of static sensor nodes randomly deployed in the target area.

a. System Environment

We assume n sensor nodes randomly distributed in the area to be monitored are continuously sensing and reporting events. These sensor nodes are static. We use S_i to indicate the i-th node, sensor nodes set S = {S_1, S_2, ... , S_n}, and the number of S is n. We make the following assumptions about the sensor nodes and the network module.

i) The sink is deployed in a region away from the sensors and we assume that the energy of the sink is infinite.
ii) Sensor nodes will be assigned a unique identifier before deployed in the sensing area.
iii) All nodes have the same computing, storage and energy capabilities.
iv) The sensor node's transmission power can be changed according to the distance from the receiver.
v) All sensor nodes are static. In addition, each sensor node knows its own location and the sink knows their location though the use of the location mechanism.

b. *Energy Consumption Module*

Our paper is using formula from W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan [7] for the communication energy consumption module, and following the formula:

$$E_T(k,d) = E_{Tx}k + E_{amp}(d)k \quad (1)$$

$$E_R(k) = E_{Rx}k \quad (2)$$

$$E_{fuse}(k) = E_{fuse}k \quad (3)$$

Formula (1) means the sensor nodes have an energy cost when transmitting data, (2) means the sensor nodes have an energy cost when receiving data, and (3) means the sensor nodes have an energy cost when the data fuses. In these three functions, where k is data packet size, E_{Tx} is energy cost for transmitting one unit data of the sensor node, E_{Rx} represents the energy cost when node receives one unit data, E_{fuse} is energy cost of the fusing the data. When the sensor nodes transmit amplification is required, so transmitting nodes have an additional $E_{amp}(d)k$ energy cost. The value of $E_{amp}(d)k$ can be determined by formula (4)

$$E_{amp}(d)k = \varepsilon_{FS}d^2 \quad (4)$$

where d is the distance between two nodes, ε_{FS} represents the amplified electric power energy cost.

c. Sensor Node and Cluster Information

Table 1 shows the sensor node information, which is used to record information about itself. Next we will introduce each field of the table. Node_ID is the identification of the node. Res_Energy is the residual energy of the node. Head_ID is the identification of the cluster head in its own cluster, if the Head_ID and Node_ID are the same, the node itself is the cluster head. Cluster_ID is the cluster number the sensor node belongs to. Next_Hop is the next sensor node to forward data to. Table 2 is the cluster table, it records information of every cluster member and including the following fields Node_ID, Res_Energy, and Cost. Node_ID is the identification of member of the node in the cluster.

TABLE I. SENSOR NODE INFORMATION TABLE

Node_ID	Res_Energy	Head_ID	Cluster_ID	Next_Hop

TABLE II. CLUSTER_TABLE

Node_ID	Res_Energy	Cost

 ISBN: 978-1-61208-203-5

TABLE III. HEAD_LIST

Node_ID	Cluster_ID	Cost

Res_Energy is the residual energy of the node in the cluster. Cost is transmission cost between two nodes. Table 3 is used to record information one neighboring cluster heads. The fields include Node_ID, Cluster_ID and Cost. Node_ID is the identification of the neighboring cluster head. Cluster_ID is the cluster number of the neighboring of cluster head. Cost is transmission cost between two cluster heads.

B. Clustering Phase

a. Clustering

In this paper, MGCRP merge neighboring nodes as far as possible in the same cluster. Before discussing the clustering step, first, we must are define the routing protocol parameters and variables.

- Rectangle Unit Block (Block)**:** This is a rectangular block. The user defined value of N divides the network into several blocks which are the same size and are not overlapping.
- Center of Block (BC)**:** After dividing the network into several grids, we will calculate the center coordinates of the grids resulting in a vector of coordinates. We assume Block_Center i (BC_i) is the i-th center coordinates of the block.
- Cluster**:** Cluster can be regarded as a set C which is includes several sensor nodes. We can represent a set C as $C = \{S_j\}, S_j \in S, j=1,2,\ldots,n$. Where j is the number of sensor nodes. We assume the Cluster_ID is i which represent the cluster number of the grid. In any cluster (C_i), if any member of the sensor nodes are not in a cluster, it is an invalid cluster, otherwise, it is valid cluster.
- Distribution**:** We define a new parameter in a valid cluster Distribution it is used to evaluate the distribution of nodes in a valid cluster. The number of nodes within a valid cluster must be closer to BC. The formula (5) is used to calculate the distribution of the cluster. Where $d(S_m, BC_i)$ is the distance between the member of the sensor nodes in cluster and the BC of the cluster and where $N(C_i)$ is the number of sensor nodes in the cluster.

After defining these parameters and variables, the following details the description of each step.

Step 1: Network Gridding

After the deployment of the sensor nodes, we will make **a grid of the** network. In this paper, we assume that the sensor nodes in the network can be arranged to an M*M area, and assume every length of block is N. The network will be divided into $(\frac{M}{N})^2$ same size blocks. Where the user defined the N value and M value is the length of the sensor network.

Step 2: Calculate Center of Grid

Formula (6) calculates the center of the grid. BC_i is the two-dimensional coordinate vector, where i=1,2,…, $(\frac{M}{N})^2$, this is used to indicate the number of the grid, and also is also the Cluster_ID. The numbering starts from the (0,0) position along the X axis towards the right, Sequenced 1, 2, …, until numbered to the right- border of the sensor network, then back to left-border of the sensor network. In this moment, shift the Y-axis direction one unit block down, then repeat the sequencing step until the grid is complete. Then use the number of grids and formula (6) to get each center of grid.

Step 3: Calculate Distribution (C_i) of Valid Cluster

In this step, we will calculate the Distribution of the valid cluster. First, we give a set VC that includes all valid clusters in the network. N(VC) expresses the number of valid clusters. We will only calculate the Distribution of clusters in the VC set. After each Distribution in each cluster has been calculated, we will start the cluster merging process.

$$Distribution(C_i) = \frac{[\sum_{S_m \in S} d(S_m, BC_i)]}{N(C_i)} \quad (5)$$

$$BC_i = \left(\left[(i-1)\%\left(\frac{M}{N}\right)+\frac{1}{2}\right]*\frac{M}{N}, \left[\left|(i-1)/\left(\frac{M}{N}\right)\right|+\frac{1}{2}\right]*\frac{M}{N}\right) \quad (6)$$

Step 4: Merge Valid Cluster

First, we choose the fewest number of nodes and the cluster with the largest Distribution value from the VC set. Assume a cluster from the VC set that meets the above conditions is C_A, where $C_A \in VC$, A=1,2,…,$(\frac{M}{N})^2$, then we will start the merge. Let the distance between S_a and S_b be minimal, where $S_a \in C_A$, $S_b \in C_B$,B=1,2,…, $(\frac{M}{N})^2$, $C_B \in VC$ and $C_B \neq C_A$. Then we add all the sensor nodes to C_B fromC_A. In other words, let all the Cluster_IDs of the sensor nodes from C_A change to C_B, and remove from the VC C_A, resulting in one less N(VC) .

Step 5: Clustering Finish

Assume the variable K is the user set up number of clusters in the network. The value of K will affect the efficiency of network, so we must decide the variable K according to the network size and number of nodes. The operation of clustering in step 4 will be repeated until N(VC)=K. After clustering finishes, the sink will send related information to the sensor node for an update.

b. Cluster Head Selection

The main task of the cluster head is to fuse data that sensor nodes sensed within a cluster, receive other cluster heads' sensed data, and, send to sink, after clustering finishes and, cluster head must be selected from each cluster. To do so, the sink will broadcast a Head_Elect Message packet to every sensor node in each cluster in the network. When a sensor node gets this packet, it will generate a random variable P between 0 and 1, where *P* is used to differentiate between the same residual energy from other sensor nodes. After sensor node got a random variable *P*, then immediately to calculate itself residual energy. The residual energy is then calculated.

The member nodes of the same cluster compare each of their residual energies according to the transmission power to obtain cost between them. Sensor nodes with the most residual energy will be selected as the cluster head. If more than one sensor nodes have the same residual energy in the same cluster then the sensor node with the larger *P* value will be selected as cluster head. After each of the cluster heads of cluster has been selected, each cluster head will send a Head_Confirm packet to

 ISBN: 978-1-61208-203-5

TABLE IV . HEAD_CONFIRM PACKET

Header	Node_ID	Cluster_ID

the sink. The packet format is shown in Table 4. Each Head_Confirm packet contains three fields, they are Header, Node_ID and Cluster_ID. The Header records the name of packet, Node_ID expresses the Node_ID of the sensor node that is the cluster head, Cluster_ID expresses the Cluster_ID of the cluster to where the cluster head belongs. After the sink receives all the Head_Confirm packets, it will consolidate the information and forward it to each cluster head allowing them to, update their Head_List table.

C. Route

After the clustering stage and the cluster head selection stage, the cluster structure has been established and is complete. The sensor nodes will start sensing and continuously monitor and then through a routing path, start to transmit data. In our paper, the transmission route can be divided into two parts: inner-cluster transmission route and outer-cluster transmission route. Inner-cluster transmission route refers to the path between cluster head and sensor nodes within the same cluster. Outer-cluster transmission route refers to the path between cluster heads.

Our proposed path selection method is mainly based on transmission cost between the sensor nodes. So we use Bellman-Ford shortest path algorithm for the route selection method. We arrange the network as a graph, and assume the sensor nodes in the network are the vertexes of graph and the transmission cost between nodes are edges of the graph. Through the Bellman-Ford algorithm we can calculate the lowest cost of each sensor node to the other.

a. Inner-Cluster Transmission Route Build

Inner-cluster transmission routes are the path between sensor nodes within the same cluster. The sink broadcasts to all the sensors nodes their minimum path cost to the cluster head in their cluster using the Bellman-Ford algorithm according to member cost in the Cluster_Table. In (7), $C(i,j)$ defines the cost between node i and node j, where $P_t(i,j)$ is the transmission power of node i to node j during transmission. After the sensor node receives the minimum cost between the node and the cluster head, it then records the next hop target in the Next_Hop field. When the node wants to transmit data, it sends data to the sensor node based on Next_Hop field. During the transmission, if the sensor node dies or cluster head changes, the Bellman-Ford algorithm is invoked to re-calculate the minimum cost path and the Next_Hop field is updated.

$$C(i,j) = P_t(i,j) \quad (7)$$

b. Outer-Cluster Transmission Route Build

The outer-cluster transmission route and inner-cluster transmission route have the same algorithm, but in the outer-cluster, the send object changes to cluster head to cluster head. The cluster head receives the data that members sent in the cluster and, then it integrates the received data and forwards it to its neighboring cluster head. According to the Cost field of Head_List and through the use of the Bellman-ford algorithm, the cluster head selects the next hop. After the calculation, the cluster head will record the transmission object. If the cluster head has been replaced, then the new cluster head will request a member to re-calculate the minimum cost between cluster heads.

c. *Cluster Head Rotation Mechanism*

The cluster head not only senses the environment but integrates the data from members of the same cluster, and transmits data to other cluster heads. In order to reduce the early death of sensor nodes, we add the cluster head rotation mechanism to distribute the energy consumption. We assume time divided into continuous periods of *T*, in the beginning *T* the sink will send a Cluster_Head Rotation Message to the sensor network. After a normal node receives this message, they will immediately send their residual energy information to the cluster head. Then the cluster head will select the node of with the most residual energy to be the new cluster head. At the same time, the cluster head will broadcast to members within cluster the new identify of the cluster head and update Head_ID of the normal node and send Head_Confirm packet to the sink. The sink will gather all the new cluster head information, consolidate, and send the data to all the cluster heads so they can update Head_List table. During *T*, the sink will repeat the above action to replace the cluster head until the energy of members within cluster is less than the energy defined by the cluster head threshold.

IV. SIMULATION AND ANALYSIS

A. Simulation Environment

This paper uses the Dev C++ simulation environment. The conditions of the sensor network and its related values by W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan [7] are shown on Table 5. Figure 1 shows the relational chart between LEACH, EBCRP and MGCRP, which shows the number of live nodes and number of rounds. Fig.2 is the relational chart between LEACH, EBCRP and MGCRP, which shows the average energy consumption and number of rounds. A round is defined by data that is transmitted to sink safely; a conclusion that is made from the average of 50 kinds of conditions.

B. Simulation Results

We can observe that MGCRP is better than LEACH and EBCRP via Figure 1 and Figure 2. The selection of the cluster head method of LEACH is random, and the clusters transmit collected data directly to the sink. So if there are several clusters which are far away from the sink, the network would die from a large consumption of energy reducing, the number of data transmissions. The selection of cluster head principle of EBCRP is better than LEACH because, it chooses the nodes which are closer to sink to be clusters. The design does not have the transmission distance limitation of clusters in LEACH, but the routing of EBCRP is a chain which is connected by nodes resulting in redundant data transmissions. In this paper, MGCRP combines nodes which are closer to others in a cluster

 ISBN: 978-1-61208-203-5

and when the nodes are distributed unevenly, it shortens the distance between nodes and the cluster head to attain power savings. To not overload any one member, we add cluster head rotation mechanism, to equally to distribute energy consumption to the members in the cluster prolonging the network lifetime.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a routing protocol which is a clustering-based routing protocol for wireless sensor networks. In this routing protocol, we grid the network and then combine these grids based on a user defined value, There are several advantages of this protocol show below. First, sensors will be allocated by high density in the same cluster no matter what the condition is. Second, we add the cluster head rotation mechanism, and it could allot workload equally to every node. And we choose Bellman-Ford algorithm, spend the minimum of cost to transmit the data to clusters. Through effective clustering, the routing protocol which we proposed could save more energy and prolong the network lifetime.

In future, we hope to add redress mechanisms in the transmission of data stage, because when the sensor nodes may be faced with in the time of passing information to a passing objects have been killed, resulting in the data cannot pass and makes the collection of good data must be discarded. When the sensing data is discarded at the same time also means that before passing the sum of data consumed by the power follow the waste, the data must be retransmitted.

ACKNOWLEDGEMENT

Authors appreciate the funding support for the research project from National Science Council, Taiwan.
Project ID NSC 100-2221-E-032-041-

TABLE V . EXPERIMENTAL PARAMETERS

Parameter	Value
Sensing range (m^2)	**(0,0)~(100,100)**
Sink location	**(50,150)**
Sensor node numbers (n)	**100**
Sensor node initial energy (E0)	**0.5 J**
E_{Tx}, E_{Rx}	**50 nJ/bit**
ε_{FS}	**10 pJ/(bit•m^2)**
E_{fuse}	**5 nJ/(bit•single)**
Data packet size	**4000 bits**
Grid length (N)	**10 m**
Cluster number (K)	**5**

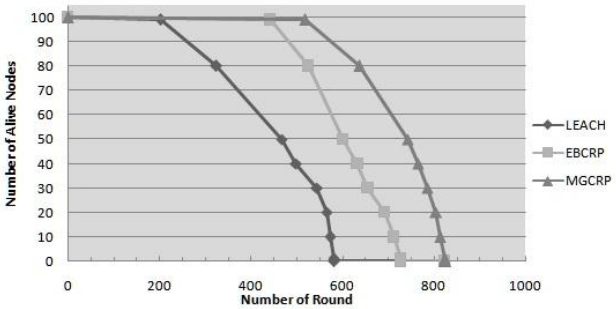


Figure 1. Relation between the number of alive nodes and number of round with different routing protocol

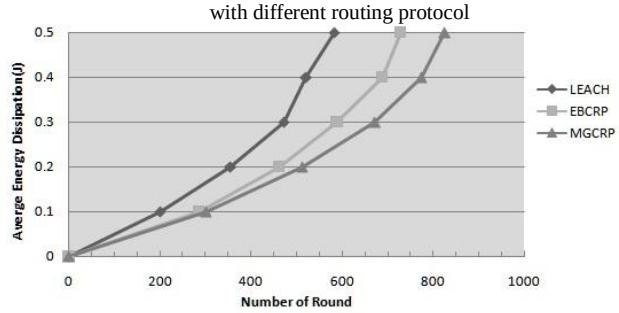


Figure 2. Relation between consumption of average energy and number of round with different routing protocol

REFERENCES

[1] I. F. Akyildiz, S. Weilian, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks, " *IEEE Communications Magazine*, vol. 40, No 8, pp. 102-114, Aug. 2002.

[2] J. N. Al-Karaki and A. E. Kamal, "Routing techniques in wireless sensor networks: a survey, " *IEEE Wireless Communications*, vol. 11, No 6, pp. 6-28, Dec. 2004.

[3] J. Qiangfeng and D. Manivannan, "Routing protocols for sensor networks," *Proceedings of First IEEE Consumer Communications and Networking Conference, 2004, CCNC 200* , pp. 93-98, Jan. 2004.

[4] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, pp. 1-10 vol. 2, Jan. 2000.

[5] X.-r. Bao, S. Zhang, D.-y. Xue, and Z.-t. Qie, "An Energy-Balanced Chain-Cluster Routing Protocol for Wireless Sensor Networks," *Proceedings of the 2010 2th International Conference on Networks Security Wireless Communications and Trusted Computing (NSWCTC)*, pp. 79-84, Apr. 2010.

[6] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Englewood Cliffs, New Jersey: Prentice–Hall, 1992.

[7] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks, " *IEEE Transactions on Wireless Communications*, vol. 1, No 4, pp. 660-670, Oct. 2002.

 ISBN: 978-1-61208-203-5

WSNs Coverage Hole Partial Recovery by Nodes' Constrained and Autonomous Movements Using Virtual α-chords

Ali Rafiei, Mehran Abolhasan, Daniel R. Franklin
University of Technology Sydney, Australia
ali.rafiei@student.uts.edu.au
{mehran.abolhasan, daniel.franklin}@uts.edu.au
Farzad Safaei
University of Wollongong, Australia
Farzad@uow.edu.au

***Abstract*—In WSNs, in order to recover from coverage holes and to mitigate their indirect/direct effects on networks' performance, different recovery strategies such as increasing proximate nodes' transmission range and/or relocation of nodes towards coverage holes seem to be appropriate solutions. Since the majority of a mobile node's energy is consumed by movement and since nodes' residual energy may be affected by damage events, node movements should be performed sparingly. Conventional nodes' information exchange in real-time applications with security and interference concerns are neither practical nor secure. Therefore, for the aforementioned scenarios, at the price of possible node collisions, disconnections, and reasonable compromises, promising distributed and autonomous node movement algorithms based on limited 1-hop neighbour knowledge are proposed. Our proposed autonomous and constrained node movement model based on a node's 1-hop perception provides a feasible and rapid recovery mechanism for large scale coverage holes in real-time and harsh environments. Our model not only maintains moving nodes' connectivity to the rest of network to some extent, but also offers emergent cooperative recovery behaviour among autonomous moving nodes. Our movement model based on virtual chords formed by nodes and their real and virtual 1-hop neighbours, not only confines node movement range, but also takes the issue of moving nodes' connectivity into account. Suitable performance metrics for partial recovery via constrained movement are introduced to compare the performance and efficiency of our model with conventional Voronoi-based movement algorithms. Results show that our proposed model's performance is comparable with Voronoi-based movement algorithms.**

***Index Terms*—Coverage holes; autonomous and constrained movements; Wireless sensor networks; virtual chord.**

I. INTRODUCTION

Due to the vast applications of wireless sensor networks (WSNs) [1][2], they are a key focus of attention for academic and industrial research. Deployed sensor nodes [3] can be used to detect fire [4], tsunamis [5], to monitor wildfire [6], earthquakes [7], habitats [8], environment [9], and active volcanoes [10]. New generations of sensors deployed and embedded in a variety of environments such as structures [11], underground [12], air (as unmanned aerial vehicles) [13], underwater [14], or on the sea surface [15] can be used to detect many events and phenomena, notify other nodes, and respond to the events. In addition to emerging WSN applications, diverse nodes' deployment [3], mobility, and movement patterns [16] offer new remedies to WSNs' challenges [17][18]. Despite continuous reduction in cost/size and increase in nodes' battery/processing power, an economically justifiable degree of redundancy in deployed nodes should be considered in order to have flexibility and robustness in node failure-prone environments with harsh conditions. Depending on application and environment, a trade-off between nodes' density [3] and mobility [19] (uncontrolled and controlled movements [20]) should taken into account if a proper level of quality of service is to be achieved.

Having severe direct/indirect effects on the networks' integrity and performance, large scale *coverage holes* caused by en masse node failures in a given area(s), should be avoided [21] and/or mitigated as much as possible with different recovery strategies. In WSNs, it is not always possible to deploy new nodes in unsupervised and harsh environments and dropping random nodes cannot guarantee desirable node formations and distributions. Although it may not be so economical, by benefiting from the redundant nature of deployed nodes, coverage holes to some extent can be repaired either by transmission power adjustment or the relocation of a selected set of currently deployed nodes (e.g., damaged area proximate nodes). Since movements consume the majority of nodes' energies, they should be moved carefully. Therefore, the amount of movements for proximate nodes known as *boundary node* (*B-nodes*) which participate in the recovery of damaged areas should be done sparingly; otherwise, nodes' energy exhaustion results in further cascaded failures.

Though for precise nodes movements a reasonable amount of message exchanges are required, in real-time scenarios with security and interference considerations, they are neither desirable nor secure. Therefore, by putting the burden of autonomous decision and more processing on individual B-nodes who directly detected the damage events, the number of exchanged messages can be kept as small as possible. Autonomous movement decision-making has the drawback

 ISBN: 978-1-61208-203-5

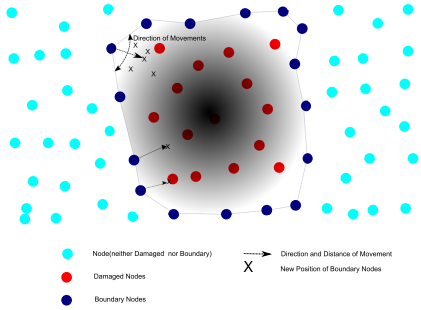


Fig. 1: Coverage Hole and Node Types

of increasing the possibility of collision and disconnection among nodes. Moreover, improvised, unconstrained, and careless movements towards damaged area(s) may cause multiple newly formed coverage holes. It should be noted, however, that for the sake of temporal coverage, a coverage hole may be virtually displaced via controlled group node movements as in [22]. Our model of autonomous and constrained node movements towards the coverage hole tries to maintain connectivity of the moving nodes with their 1-hop immediate neighbours. These autonomous movements in each of the moving nodes are inferred solely from the limited knowledge of node's 1-hop neighbours before the *damage event* as well as the perceived 1-hop neighbors' status change after damage event without *any* additional message exchanges. In our model each boundary node forms a α-*virtual chord* through its selected real and virtual neighbours (the other endpoint of the given chord) with the length of $\alpha \cdot 2 \cdot R_c \leq 2 \cdot R_c$ $(0 \leq \alpha \leq 1)$ (Fig. 2). Each virtual chord's endpoints (i.e., node's real and virtual neighbours) lie on the circumference of the two distinct circles with equal radius of R_c. One of these circles is considered a *valid circle* if it is closer to the damaged area. With α-virtual chord node movement, relocated B-node from s to s' maintains the connectivity with its real and virtual neighbours n and n', provided its real neighbour n is not a moving B-node. In our proposed model, not only an ensemble nodes' emergent cooperative movement behaviour [23] is manifested, but also group mobility and cooperative behaviour of moving nodes can be changed by using different values of α. So by changing α, the direction of moving nodes towards the coverage hole changes to the direction of nodes circulating around it. The former is suitable for the case of hole recovery while the latter is geared to prevent cascaded failure and failure expansion around damaged area as lower residual energy B-nodes can be replaced with other moving nodes with higher energy as a result of nodes constrained circular movements (Fig. 3). These types of different group mobility behaviours can be implemented by collection of nodes' autonomous movements via local decisions made based on simple nodes' geometrical and statistical features.

To our best knowledge, very few works considered partial recovery of large scale coverage hole via autonomous constrained node movements for time-sensitive scenarios with security consideration. Moreover, few works used damaged nodes' statistical and geometrical features as the landmarks in nodes' local and autonomous decision making processes. We have also defined proper performance metrics in order to compare the performance and efficiency of our proposed model with conventional Voronoi-based movement models (VOR and MinMax) [24][18]. In Section II, we present current work on nodes movement. In Section III, our model and assumptions are introduced. In Section IV, our proposed performance metric are briefly discussed, and finally, in Sections V and IV, result, conclusion, and future work are respectively presented.

II. Related Work

Mobility in wireless sensor networks is a double edge sword; on one hand, undesirable and uncontrolled mobility causes coverage holes and topological instability, while on the other hand, coverage hole(s) can be repaired by controlled mobility and movement of nodes [20][17]. Thus node relocations [18] are important in enhancing networks coverage and connectivity [25], by offering *temporal coverage* in addition to *spatial coverage* [26] for an area in which the number of nodes is not sufficient to cover it all time. Thus via controlled mobility, a trade-off between the number of required deployed nodes and the required coverage of the given area can be reached [19]. Controlled mobility not only is able to repair the coverage hole but also it can correct irregularities of uncontrolled mobility [20]. After deployment, especially in hostile and hazardous environments, it is almost impossible to have centralised control over sensors. Thus, in such case in order to repair coverage holes, nodes not only should be able to decide autonomously on their movements but also they should not exhaust their energy as majority of nodes' residual energy would be consumed by their movements. There are a variety of relocation, movement and deployment model in the literature [18][24][27][28][29] [30][31][32][33] which mainly aim to keep network coverage, balance node deployments, and repair small coverage holes due to improper node deployments, single or random node failure.

Movement algorithms can be divided into (virtual) radial [33] and angular [32] *force-based*, *flip-based* [28] and *Voronoi-based* [24] movement algorithms. Movement based on virtual potential repulsion and attraction [33] between pairs of nodes and the movement of nodes as the result of aggregation of these forces are inspired by physical laws of nature. Virtual angular force [32] tries to connect the partitions and parts of network by using collaborative movement of mobile nodes applying on the angle of moving nodes.

In order to exert proper levels of virtual repulsion and attraction, nodes should be globally aware of the their targeted density. Since the movement algorithm is applied to all nodes, movement contains oscillation due to mutual interaction of nodes; consequently, an unnecessary amount of nodes' energy is consumed. In flip-based movement algorithm [28], the given area is divided into regions and a head node is elected for each region. In the case of head failure and unbalance number of nodes, nodes from the neighbour regions would flip into the given region. In flip-based movement,

 ISBN: 978-1-61208-203-5

the head node for each region should be selected, which requires message exchange among nodes. Since movement is confined to neighbouring regions, the recovery of large scale coverage hole may consists of many iterations of nodes flipping into their neighbouring regions with an agreed-upon granularity. So flip-based movement algorithms are expected to be inefficient for real-time scenarios with large scale holes. In Voronoi-based movement algorithms, [24], the area is decomposed into Voronoi diagrams [34] depending on the deployment and distribution of nodes. If a node fails or part of area is not covered by the sensor network, nodes move with regard to their Voronoi vertices to compensate for *void area(s)*. Voronoi-based movement most often is required to have global knowledge to form Voronoi diagrams. Voronoi-based movement algorithms are not geared for large scale coverage holes as they result in newly formed small coverage holes. They also suffer from oscillation and consequently energy exhaustion if recovery is performed in an iterative style. Complex and centralised node movements and even distributed algorithms (with pre-computed movements) have a good energy management, however, they are not efficient for real-time scenarios as they suffer from unacceptable delay, particularly under very fast-changing conditions. So diffused information and nodes' notifications are not valid and already obsolete for the decision making process.

III. Methods and Assumptions

A. Sensor Model and Node Types

Homogeneous sensor nodes are modelled based on the unit disk graph (UDG) [35] and are bidirectionally connected if they are within each other's ranges. Nodes are randomly deployed with uniform distribution in a rectangular area of $[x_{min}, x_{max}] \times [y_{min}, y_{max}]$. To avoid unnecessary complexity, it is assumed that transmission range (R_c) and sensing range (R_s) are equal. Although no central coordination is required and a local coordination system is applicable in our model, sensors' locations may be known by GPS or any other localisation methods [36]. Sensor nodes are classified into *damaged* nodes (*D-nodes*) if they reside inside the *damaged area* (*D-area*); otherwise, they are considered as *undamaged* nodes (*U-nodes*). Those proximate U-nodes to D-area which directly detect the *damage event* (*D-event*) within their ranges are further classified into *boundary nodes* (*B-nodes*). B-nodes detect the D-event as they sense any significant changes within their ranges such as signal loss or disconnection due to the failure of their neighbours. It should be noted that noise, false alarms, or transient, periodic, and frequent failure of nodes, and link instability are excluded in our model.

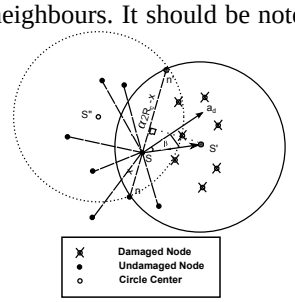


Fig. 2: B-Node, its real, virtual neighbors and virtual chord

B. Coverage hole

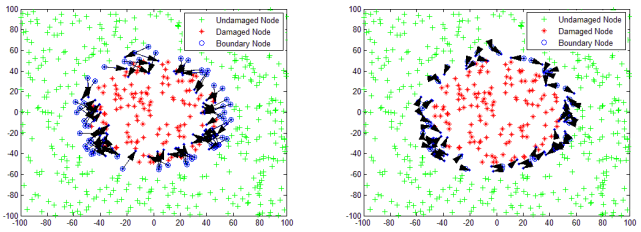
Coverage hole are modelled in different forms in the literature [22][37][38]. Similar to [39][40], coverage holes are modelled as a circle of radius r_{Hole} with the centre at x_{Hole}, y_{Hole}. Since each B-node autonomously perceives the D-event and its damaged neighbours, a *margin* of B-nodes (*MB-nodes*) are formed around the D-area. Thus, to benefit from WSNs redundancy and to reduce possibility of interference and collision, a set of the B-nodes defined as *selected B-nodes* (*SB-nodes*) [39][40] are selected which may partake in a possible recovery process by moving towards the *region of interest* (*ROI*). SB-nodes may be selected by a distributed algorithm or centrally selected based on the agreed criteria. Similar to [39][40], B-nodes are selected in a distributed fashion based on B-node's 1-hop geometrical and statistical features.

C. Selected Neighbor Nodes

B-node's neighbours can be classified into D-nodes or U-node depending on their location relative to the coverage hole. Based on the type of B-node's neighbours, they can be defined as the *undamaged neighbour nodes* (*UN-nodes*) or *damaged neighbour nodes* (*DN-nodes*. At the time of the D-event, each B-node's distances to both sets of its UN-nodes and DN-nodes as well as their degrees of connectivity are used as *landmarks* in decision making processes. Therefore, if a B-node selects a set of its UN-node(s) via some selection algorithms, those UN-nodes are considered as *selected undamaged neighbour nodes*(real neighbours). *Virtual selected undamaged neighbour nodes* are the fictitious B-nodes' neighbours (virtual neighbours) which are connected to B-node's UN-nodes via *virtual chords* defined as α-chord with the length of $\alpha \cdot 2 \cdot R_c \leq 2 \cdot R_c$ ($0 \leq \alpha \leq 1$) (Fig. 2). Three neighbour node selection algorithms, namely *closest neighbour*, *random neighbour*, and β*-angle* are presented in Algorithm 1.

In the closest neighbour algorithm, a B-node's closest 1-hop neighbour is selected, while in the random select algorithm one of the B-node's 1-hop neighbours is randomly selected. In the β-angle algorithm, for each B-node and the undamaged node in its neighbour set, set of angles can be formed between normal direction of the virtual chords (Fig. 2) and distance vector from the B-node to its D-nodes centre of mass. B-node's neighbour whose aforementioned angle is closer to β than any other of B-nodes's 1-hop undamaged neighbours are considered as the selected undamaged neighbour and should be unique. If more than one undamaged neighbour can be selected based on the mentioned conditions, only one of them should be randomly chosen as the selected undamaged neighbour. In finding the centre of mass of B-nodes' undamaged and damaged neighbours, if the neighbours' degrees of connectivity are taken into account (Algorithm 1) they can be considered as weighted, β-angle algorithms with $w = 1$; otherwise the are called β-angle with $w = 0$.

 ISBN: 978-1-61208-203-5



(a) $\alpha = 0$ (b) $\alpha = 1$

Fig. 3: Chord Movement Algorithm (R_c=15, N=600, β=0)

D. Movement Model

Movement algorithms can be divided into following states: 1) undamaged neighbour nodes of moving B-nodes are selected based on criteria presented in Algorithm 1; 2) with regard to the suitable virtual chord parameters α and R_c, the location of B-nodes' virtual neighbours are obtained for each B-node; 3) new locations of moving B-nodes computed by selecting one of two circles which pass through the endpoints of the virtual chord of each B-node (Algorithm 2). The selected circle is defined as the *valid circle* through which the chord is obtained. The Valid circle is the circle with its centre closer to damage area; 4) the B-node then moves to the centre of the valid circle with probability p (uniform distribution) and q otherwise, such that $p + q = 1$. Here we assumed $p = 1$ in which all B-nodes move towards the coverage hole. It should be noted that connectivity of B-nodes to its neighbours can not be fully guaranteed. This is because after the damage event, B-nodes are not able to distinguish if their undamaged neighbours are moving B-nodes or or not. As an example, Fig. 3 shows how changing parameter α affects B-nodes' collective movement behaviour in our coverage hole recovery model. β-angle with $\alpha = 0/1$ in Fig. 3 show the direction of moving B-nodes towards/around coverage hole.

IV. Performance Metrics

We have compared our proposed movement algorithm with the two Voronoi-based movement algorithms (VOR-MinMax) [24] via three types of proposed performance metrics. In Voronoi-based algorithms, B-nodes were selected similarly to our previous work [40]. In modelling Voronoi movement algorithms, we have also considered the problem of nodes with out-of-area and infinite Voronoi vertices. The proposed performance metrics are classified below:

Coverage-based metrics: We define *percentage of recovery* as the percentage of recovered networks' 1-coverage after the recovery process. In other words, the metric shows by using the given movement algorithm what percentage of lost 1-coverage is recovered in the network.

Connectivity-based metrics: We define *percentage of connectivity* as the percentage of moving B-nodes which are directly connected to rest of network (those nodes which did not participate in the recovery process) with at least one link over the total number of moving B-nodes. This performance metric shows the effect of movement algorithms on the connectivity of moving nodes and how many of the moving B-nodes are still directly connected to the rest of the network after their movements.

Algorithm 1: Nodes' neighbors selection Algorithms

Input:
s_i^b : B-node i ($i = 1, \cdots, m$), $N_{s_i}^h$: s_i^b's h-hop neighbours
$N_{s_i}^{h_u}$: h-hop U-node neighbours of s_i^b
$N_{s_i}^{h_d}$: h-hop D-node neighbours of s_i^b
$\overrightarrow{X}_{s_i}^{s_j^{h_u}}$ distance vector from s_i to s_j (j in $N_{s_i}^{h_u}$)
$\overrightarrow{X}_{s_i}^{s_j^{h_d}}$ distance vector from s_i to s_j (j in $N_{s_i}^{h_d}$)
$\beta - angle$: angle parameter
$d_{S_j}^{h_u}$ degree of s_j (j from $N_{s_i}^{h_d}$)
$d_{S_j}^{h_d}$ degree of s_j (j from $N_{s_i}^{h_u}$)
Output: Set of selected h-hop neighbour $s_j^{h_u} s_i^b$

case *Closest* if closest neighbour selected
  foreach *B-Node* s_i^b **do**
    Find $N_{s_i}^{h_u}$
    foreach *h-hop UN-nodes* $s_j^{h_u}$ **do**
      Calculate $\overrightarrow{X}_{s_i}^{s_j^{h_u}}$
    Calculate $arg_j Min \left(\left| \overrightarrow{X}_{s_i}^{s_j^{h_u}} \right| \right)$
case *Random* if Random neighbor Selected
  foreach *B-Node* s_i^b **do**
    Find $N_{s_i}^{h_u}$
    Calculate $arg_j Random \left(N_{s_i}^{h_u} \right)$
case β*-angle* if β-angle is Selected
  foreach *B-Nodes* s_i^b **do**
    Find $N_{s_i}^{h_u}$ and $N_{s_i}^{h_d}$
    foreach *h-hop(DN-node* $s_j^{h_d}$, *UN-node* $s_j^{h_u}$) **do**
      Find $d_{S_j}^{h_d}$, $d_{S_j}^{h_u}$
      Calculate $\overrightarrow{X}_{s_i}^{s_j^{h_d}}$,$\overrightarrow{X}_{s_i}^{s_j^{h_u}}$
    Calculate $\overrightarrow{X}_{CM_{si}}^{h_d} = \frac{\sum (\overrightarrow{X}_{s_i}^{s_j^{h_d}}) \cdot d_{S_j}^{h_d}}{\sum d_{S_j}^{h_d}}$
    Calculate $\overrightarrow{X}_{CM_{s_i}}^{h_u} = \frac{\sum (\overrightarrow{X}_{s_i}^{s_j^{h_u}}) \cdot d_{S_j}^{h_u}}{\sum d_{S_j}^{h_u}}$
    foreach *h-hop UN-node* $s_j^{h_u}$ **do**
      Calculate $\angle \gamma_{s_i}^{h_u S_j} =$
      $\angle \left(\overrightarrow{X}_{CM_{si}}^{h_d}, \overrightarrow{X}_{s_i}^{s_j^{h_u}} \right) - \angle \beta$
    Calculate $arg_j Min \left(\left| cos(\angle \gamma_{s_i}^{h_u S_j}) \right| \right)$

Algorithm 2: Formation of Chord Algorithm

Input:
s_i^b : B-node i $(i = 1, \cdots, m)$, τ : threshold
$s_j^{h_{u_{s_i^b}}}$: Selected h-hop U-node neighbour s_j
α-chord parameter, R_c transmission Range
$N_{s_i}^{h_u}$: h-hop U-node neighbours of s_i^b
$N_{s_i}^{h_d}$: h-hop D-node neighbours of s_i^b
Output:
B-nodes s_i^b's new location (coordinates), $s'^b_i(x, y)$

foreach *B-node* s_i^b **do**
  Find s_i^b's current location (coordinates) of $s_i^b(x, y)$
  Find $CM_i^{h_{u(x,y)}}$: s_i^b's h-hop UN-nodes' center of mass
  Find $CM_i^{h_{d(x,y)}}$: s_i^b's h-hop DN-nodes' center of mass
  Calculate $chord - \alpha_i$, virtual node $s'^{h_{u_{s_i^b}}}_j$ from α_i and R_c
  Find $C_{\alpha_i}^{(x,y)_{k,k'}}$ (circle center(s)) of $chord_{\alpha_i}$
  foreach $chord_{\alpha_i}$ *and* $C_{\alpha_i}^{(x,y)_{k,k'}}$ **do**
    if $\left\| C_{\alpha_i}^{(x,y)_k} - CM_i^{h_{d(x,y)}} \right\| < \left\| C_{\alpha_i}^{(x,y)_{k'}} - CM_i^{h_{d(x,y)}} \right\|$ **then**
      $C_{Valid_{\alpha_i}}^{(x,y)} = C_{\alpha_i}^{(x,y)_k}$
    else if $\left\| C_{\alpha_i}^{(x,y)_k} - CM_i^{h_{d(x,y)}} \right\| > \left\| C_{\alpha_i}^{(x,y)_{k'}} - CM_i^{h_{d(x,y)}} \right\|$ **then**
      $C_{Valid_{\alpha_i}}^{(x,y)} = C_{\alpha_i}^{(x,y)_{k'}}$
    else if $\left\| C_{\alpha_i}^{(x,y)_k} - CM_i^{h_{d(x,y)}} \right\| = \left\| C_{\alpha_i}^{(x,y)_{k'}} - CM_i^{h_{d(x,y)}} \right\|$ **then**
      Calculate rand $p \sim U[0, 1]$
      if $p > \tau$ **then**
        $C_{Valid_{\alpha_i}}^{(x,y)} = C_{\alpha_i}^{(x,y)_k}$
      else
        $p < \tau$
        $C_{Valid_{\alpha_i}}^{(x,y)} = C_{\alpha_i}^{(x,y)_{k'}}$
  $s'^b_i(x, y) = C_{Valid_{\alpha_i}}^{(x,y)}$

Distance-based metrics: We define *average movement* as the ratio of total amount of movement to the number of participating nodes in recovery process. Average movement can be used with other metrics to better understand the behaviour of movement algorithms in coverage hole recovery process.

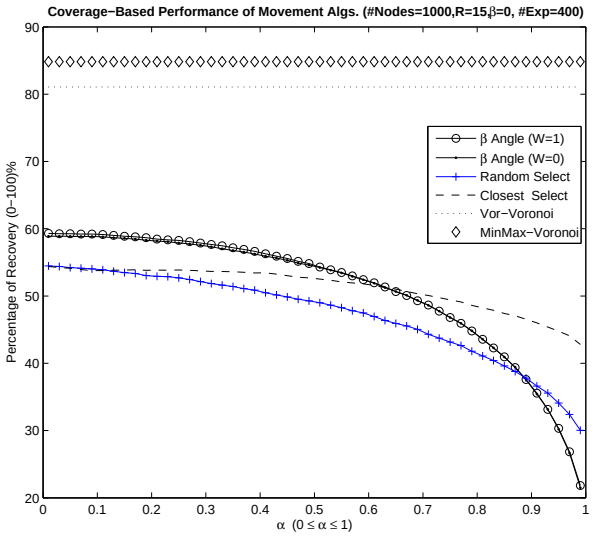


Fig. 4: Percentage of Recovery

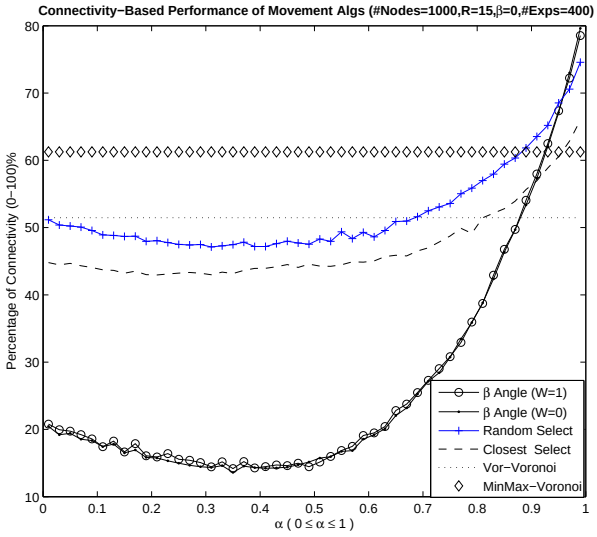


Fig. 5: Percentage of Connectivity

V. Results

Using Matlab, N=1000 nodes with communication and sensing range of 15 ($R_c = R_s = 15$ m) are uniformly deployed with random distribution in a rectangular area of $[-100, 100] \times [-100, 100]$. Similarly to [39][40], coverage holes are modelled as circles with radius $r_{Hole} = 50$ m located at $(x_{Hole}, y_{Hole}) = (0, 0)$. The experiment was repeated $\#Exp = 400$ times for all movement algorithms. Chord parameter (α) is continuously changed from 0 to 1 to examine its effect in the performance and node collective behaviour of the proposed movement algorithms. Results with error bars (97.5% confidence intervals) are not included here due to space limit (Figs. 4-6).

Performance metrics of movement algorithms are also shown in Table I. With regard to Figs. 4-6, as α continuously

 ISBN: 978-1-61208-203-5

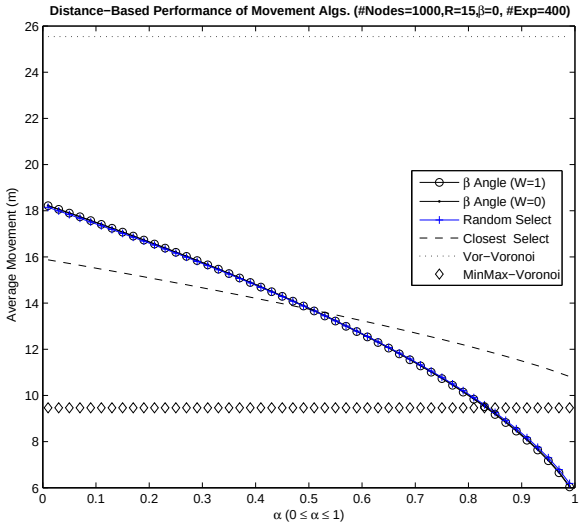


Fig. 6: Average Movement

changes from 0 to 1, the percentage of recovery and nodes' average movements of virtual chord movement algorithms decreased but at the same time their percentages of connectivity increased, which shows that B-nodes' collective behaviour and direction of movements shift gradually from moving towards to circulating around coverage hole. Each of these collective mobility behaviours can be used for different purposes. In our model, α can be chosen in such a way as to achieve proper percentage of connectivity, percentage of recovery with given amount of nodes' movement.

Results from Figs. 4-6 and Table I show that although proposed chord movement algorithm is autonomous and require few or no message exchange, its performance is comparable to Voronoi-based movements.

In the real-time scenarios with security and interference concerns, using Voronoi-based algorithms requires global knowledge of the network. So even if higher coverage and connectivity is offered, in these scenarios, they not practical. It should be noted that performance of our proposed model would change with regard to other network parameters such as network node density, node range, and coverage hole radius, node deployment distribution, etc. Therefore, their effects should be examined in more detail.

VI. Conclusion and Future Work

A new autonomous and constrained node movement model is proposed to partially/wholly recover large scale coverage holes in real-time scenarios with interference and security consideration. Our proposed model of autonomous decision making is based on the available 1-hop knowledge at the time of the damage event. By introducing the concept of α-chords, our proposed model not only taken the connectivity of moving nodes into account, but it also shows an emergent cooperative recovery behaviour. To compare our proposed model with conventional Voronoi-based algorithms, suitable performance metrics were introduced.

Algs.	α	**Recovery(%)**	**Connectivity(%)**	**Avg. Mov.(m)**
β-angle (w=1)	0	59.3000	20.5821	18.2413
	0.25	58.1561	14.9848	16.1297
	0.50	54.3100	15.7554	13.6721
	0.75	46.3983	32.1165	10.5888
	1.0	21.8333	84.2525	5.6323
β-angle(w=0)	0	58.8752	20.4436	18.2227
	0.25	57.8314	14.4888	16.1230
	0.50	54.2239	15.7458	13.6797
	0.75	46.4474	31.7605	10.6149
	1.0	21.5037	84.0388	5.6850
Closest	0	54.3857	43.9619	15.7489
	0.25	53.8395	42.9475	14.7811
	0.50	52.5149	43.9536	13.7172
	0.75	49.2130	48.7582	12.5073
	1.0	42.8042	67.2597	11.0456
Random	0	54.4647	52.0741	18.1173
	0.25	52.5196	49.3625	16.0505
	0.50	49.0044	49.2323	13.6488
	0.75	42.9059	55.0499	10.6438
	1.0	30.0398	77.9009	5.8566
Vor	0	81.0696	51.3128	25.5432
	0.25	81.0696	51.3128	25.5432
	0.50	81.0696	51.3128	25.5432
	0.75	81.0696	51.3128	25.5432
	1.0	81.0696	51.3128	25.5432
MinMax	0	84.8375	61.5663	9.4616
	0.25	84.8375	61.5663	9.4616
	0.50	84.8375	61.5663	9.4616
	0.75	84.8375	61.5663	9.4616
	1.0	84.8375	61.5663	9.4616

TABLE I: Performances of Movement Algorithms

As future work, new autonomous constrained node movements models can be defined. The issue of trade-off between nodes' amount of exchanged information and degree of node autonomy can be investigated. The problem of nodes' connectivity and collisions should also be addressed in more details in future autonomous models.

In order to show the effects of a coverage hole on its proximate nodes, node residual energy models should be included in recovery models. Undesirable secondary effects of imprudent node movements such as formation of new coverage holes should be examined more comprehensively. Probabilistic autonomous prediction of nodes' neighbours status without exchanging any additional messages to achieve emergent cooperative behaviour via autonomous nodes is also expected to be an interesting future work. New models of one-time autonomous node movements instead of iterative nodes' movements can be considered to reduce the problem newly formed coverage holes, oscillation, and energy in the network.

VII. Acknowledgment

This research was supported by the Australian Research Council (ARC) discovery research grant No. DP0879507.

References

[1] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Computer Networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[2] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, vol. 38, no. 4, pp. 393–422, 2002.

 ISBN: 978-1-61208-203-5

[3] C. Xiao, Y. Peng, and M. Yu, "The deployment method and movement control strategy in mobile wireless sensor networks," in *International Symposium on Computer Science and Computational Technology, ISCSCT 2008*, vol. 2, Dec. 2008, pp. 520–523.

[4] A. Khadivi and M. Hasler, "Fire detection and localization using wireless sensor networks," in *Sensor Applications, Experimentation, and Logistics*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, N. Komninos, O. Akan, P. Bellavista, J. Cao, F. Dressler, D. Ferrari, M. Gerla, H. Kobayashi, S. Palazzo, S. Sahni, X. S. Shen, M. Stan, J. Xiaohua, A. Zomaya, and G. Coulson, Eds. Springer Berlin Heidelberg, 2010, vol. 29, pp. 16–26.

[5] S. Bhima, A. Gogada, and R. Garimella, "A tsunami warning system employing level controlled gossiping in wireless sensor networks," in *Proceedings of the 4th international conference on Distributed computing and internet technology*, ser. ICDCIT 2007. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 306–313.

[6] Y. Li, Z. Wang, and Y. Song, "Wireless sensor network design for wildfire monitoring," in *The Sixth World Congress on Intelligent Control and Automation, WCICA 2006*, vol. 1, 2006, pp. 109 –113.

[7] M. Suzuki, S. Saruwatari, N. Kurata, and H. Morikawa, "A high-density earthquake monitoring system using wireless sensor networks," in *Proceedings of the 5th international conference on Embedded networked sensor systems*, ser. SenSys 2007. New York, NY, USA: ACM, 2007, pp. 373–374.

[8] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, "Wireless sensor networks for habitat monitoring," in *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*, ser. WSNA 2002. New York, NY, USA: ACM, 2002, pp. 88–97.

[9] T. Arici and Y. Altunbasak, "Adaptive sensing for environment monitoring using wireless sensor networks," in *Wireless Communications and Networking Conference, 2004. WCNC. 2004 IEEE*, vol. 4, March 2004, pp. 2347–2352 Vol.4.

[10] G. Werner-Allen, K. Lorincz, M. Ruiz, O. Marcillo, J. Johnson, J. Lees, and M. Welsh, "Deploying a wireless sensor network on an active volcano," *Internet Computing, IEEE*, vol. 10, no. 2, pp. 18–25, March-April 2006.

[11] S. Kim, S. Pakzad, D. Culler, J. Demmel, G. Fenves, S. Glaser, and M. Turon, "Health monitoring of civil infrastructures using wireless sensor networks," in *Proceedings of the 6th international conference on Information processing in sensor networks*, ser. IPSN 2007. New York, NY, USA: ACM, 2007, pp. 254–263.

[12] M. Li and Y. Liu, "Underground structure monitoring with wireless sensor networks," in *Proceedings of the 6th international conference on Information processing in sensor networks*, ser. IPSN 2007. New York, NY, USA: ACM, 2007, pp. 69–78.

[13] K. Daniel, S. Rohde, and C. Wietfeld, "Leveraging public wireless communication infrastructures for uav-based sensor networks," in *2010 IEEE International Conference on Technologies for Homeland Security (HST)*, Nov. 2010, pp. 179–184.

[14] I. F. Akyildiz, D. Pompili, and T. Melodia, "Underwater acoustic sensor networks: research challenges," *Ad Hoc Networks*, vol. 3, no. 3, pp. 257–279, 2005.

[15] M. Jiang, Z. Guo, F. Hong, Y. Ma, and H. Luo, "Oceansense: A practical wireless sensor network on the surface of the sea," in *IEEE International Conference on Pervasive Computing and Communications, 2009. PerCom 2009*, March 2009, pp. 1–5.

[16] R. R. Roy, *Handbook of mobile ad hoc networks for mobility models*. New York, NY: Springer, 2011.

[17] B. Liu, P. Brass, O. Dousse, P. Nain, and D. Towsley, "Mobility improves coverage of sensor networks," in *Proceedings of the 6th ACM international symposium on Mobile ad hoc networking and computing*, ser. MobiHoc 2005. New York, NY, USA: ACM, 2005, pp. 300–308.

[18] B. Wang, H. B. Lim, and D. Ma, "A survey of movement strategies for improving network coverage in wireless sensor networks," *Computer Communications*, vol. 32, no. 1314, pp. 1427–1436, 2009.

[19] W. W. V. Srinivasan and K.-C. Chua, "Trade-offs between mobility and density for coverage in wireless sensor networks," in *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*, ser. MobiCom 2007. New York, NY, USA: ACM, 2007, pp. 39–50.

[20] J. Luo, D. Wang, and Q. Zhang, "Double mobility: coverage of the sea surface with mobile sensor networks," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 13, pp. 52–55, June 2009.

[21] F. Yu, E. lee, Y. Choi, S. Park, D. Lee, Y. tian, and S.-H. Kim, "A modeling for hole problem in wireless sensor networks," in *Proceedings of the 2007 international conference on Wireless communications and mobile computing*, ser. IWCMC 2007. New York, NY, USA: ACM, 2007, pp. 370–375.

[22] C.-Y. Chang, W.-C. Chu, C.-Y. Lin, and C.-F. Cheng, "Energy-balanced hole-movement mechanism for temporal full-coverage in mobile wsns," in *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference*, ser. IWCMC 2010. New York, NY, USA: ACM, 2010, pp. 89–93.

[23] J. E. Doran, S. Franklin, N. R. Jennings, and T. J. Norman, "On cooperation in multi-agent systems," *Knowl. Eng. Rev.*, vol. 12, pp. 309–314, September 1997.

[24] M. Argany, M. Mostafavi, and F. Karimipour, "Voronoi-based approaches for geosensor networks coverage determination and optimisation: A survey," in *2010 International Symposium on Voronoi Diagrams in Science and Engineering (ISVD)*, June 2010, pp. 115–123.

[25] A. Ghosh and S. K. Das, "Coverage and connectivity issues in wireless sensor networks: A survey," *Pervasive and Mobile Computing*, vol. 4, no. 3, pp. 303–334, 2008.

[26] C. Liu and G. Cao, "Spatial-temporal coverage optimization in wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 4, pp. 465–478, April 2011.

[27] M. Younis and K. Akkaya, "Strategies and techniques for node placement in wireless sensor networks: A survey," *Ad Hoc Networks*, vol. 6, no. 4, pp. 621–655, 2008.

[28] S. Chellappan, X. Bai, B. Ma, D. Xuan, and C. Xu, "Mobility limited flip-based sensor networks deployment," *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no. 2, pp. 199–211, Feb. 2007.

[29] A. Sekhar, B. Manoj, and C. Murthy, "Dynamic coverage maintenance algorithms for sensor networks with limited mobility," in *Third IEEE International Conference on Pervasive Computing and Communications, PerCom 2005*, March 2005, pp. 51–60.

[30] X. Bai, S. Li, and J. Xu, "Mobile sensor deployment optimization for k-coverage in wireless sensor networks with a limited mobility model," *IETE Technical Review*, vol. 27, no. 2, p. 124, 2010.

[31] S. Poduri and G. Sukhatme, "Constrained coverage for mobile sensor networks," in *2004 IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA 2004*, vol. 1, April-1 May 2004, pp. 165 – 171 Vol.1.

[32] A. Casteigts, J. Albert, S. Chaumette, A. Nayak, and I. Stojmenovic, "Biconnecting a network of mobile robots using virtual angular forces," in *2010 IEEE 72nd Vehicular Technology Conference Fall (VTC 2010-Fall)*, Sept. 2010, pp. 1–5.

[33] Y. Zou and K. Chakrabarty, "Sensor deployment and target localization based on virtual forces," in *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, vol. 2, March-3 April 2003, pp. 1293–1303 vol.2.

[34] F. Aurenhammer, "Voronoi diagramsa survey of a fundamental geometric data structure," *ACM Comput. Surv.*, vol. 23, pp. 345–405, September 1991.

[35] S. Schmid and R. Wattenhofer, "Algorithmic models for sensor networks," in *20th International Parallel and Distributed Processing Symposium, IPDPS 2006*, April 2006.

[36] G. Mao, B. Fidan, and B. D. Anderson, "Wireless sensor network localization techniques," *Computer Networks*, vol. 51, no. 10, pp. 2529 – 2553, 2007.

[37] Z. Jiang, J. Wu, A. Agah, and B. Lu, "Topology control for secured coverage in wireless sensor networks," in *IEEE International Conference on Mobile Adhoc and Sensor Systems, (MASS) 2007*, Oct. 2007, pp. 1–6.

[38] G. Dini, M. Pelagatti, and I. M. Savino, "An algorithm for reconnecting wireless sensor network partitions," in *Proceedings of the 5th European conference on Wireless sensor networks*, ser. EWSN 2008. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 253–267.

[39] A. Rafiei, "Boundary node selection algorithms by simple geometrical properties in wsns," in *Fifth Asia Modelling Symposium (AMS), 2011*, May 2011, pp. 221–226.

[40] A. Rafiei, M. Abolhasan, D. Franklin, and F. Safaei, "Boundary node selection algorithms in wsns," in *The 36th IEEE Conference on Local Computer Networks (LCN), 2011*, October 2011, pp. 255–258.

 ISBN: 978-1-61208-203-5

A Technical Comparison Between Data Rate Enhancement Options in Radio Communications Networks

Cristian Androne
Communications Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
cristian.androne@com.utcluj.ro

Tudor Palade
Communications Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
tudor.palade@com.utcluj.ro

***Abstract*—This paper makes a comparative study between two newly emerged technologies in the radio communications domain: on the one hand, the small cells networks, designed to be implemented in the existing macrocellular networks, with the goal of enhancing the coverage area and the capacity of the whole network, and the 60 GHz wireless local area networks on the other hand. This latter technology is developed in order to offer high data rates taking advantage of the license free spectrum available around the 60 GHz frequency. The paper highlights the main common and disjoint aspects of both technologies and offers some implementation options.**

Keywords-60 GHz WLAN; applications; comparison; coverage;, femtocell; small cell.

I. INTRODUCTION

Due to the ever-increasing demands of today's end users, the service providers need to come up with solutions that match these requirements. The main necessities are in terms of the data transfer rates which need to be at higher levels in order for the operators to offer the desired services in radio communication networks.

Therefore, the two main fields of operation that attracted most users, i.e., cellular mobile communications and wireless local area networks, respectively, need to be enhanced in order to become viable solutions for the end users. Regarding the mobile cellular domain, new standards have been introduced, which can offer besides voice services, also data services at comfortable rates. Here, we speak of standards like HSPA/HSPA+ or LTE offered by 3GPP [1], or WiMAX offered by IEEE [2]. Even with this important enhancement regarding capacity and throughputs of the networks, services are not sufficient, especially in indoor environments, where very often the radio coverage is poor. Recent studies have shown that in cellular networks, about 60% of all voice calls and 90% of all data services, take place in indoor environments [3]. That is why it is extremely important to have a good coverage in these regions. Several recent papers present the difficulties encountered, by the traditional approach, in assuring a good indoor coverage [3],[4]. The issues relate especially to dense urban areas where it is very costly to obtain a good indoor coverage due to the geometry of the environment. Also, the capacity of the network is a sensitive problem, given the fact that using a strictly macrocellular approach, a large number of base stations would be needed, rising once again the costs. Additionally, the planning and optimization of the network would be hard to manage.

As a possible solution to these problems, the femtocell concept was developed and implemented. It is mainly designed to enhance both coverage and capacity of the traditional macrocellular network. Femtocells, also known as Home Base Stations, represent cellular network access points, which have the role of connecting the users to the operators network. The link to the macrocellular network is realized through a backhaul IP connection.

A Femtocell Access Point (FAP) is similar in concept to the wireless access point used in wireless local area networks, and it is designed to be implemented by the user. It has a low transmit power of maximum 250 mW [5], in case it is used for the residential environment. The number of active users is limited in this case, and can be up to 5 [6]. Given the fact that this equipment has a reduced transmit power, it can be implemented with a much larger density than macrocell base stations. Thus, due to the high deployment frequency, previous results show an enhanced spectral efficiency [4].

In the local area networks domain, the high density of equipment and users operating in the unlicensed ISM band has forced standardization bodies to search for alternatives to the current implementations. A possible solution is considered the implementation of the WLAN concept in the 60 GHZ frequency band. The 60 GHz millimeter wave technology is relatively new on the market and hopes to fulfill the needs of users for gigabit-scale traffic. The strong interest in the 57 – 66 GHz frequency band [7] is shown by the recent industrial and standard development efforts made by international standardization bodies like ECMA TC48, IEEE 802.15.3c and the proposed IEEE 802.11 VHT60 Task Group [8].

The high interest is due to the large bandwidth which is unmatched in any of the lower frequency bands [9]. Figure 1 shows the available spectrum for indoor wireless communications around the world. The fact that this band is unlicensed and largely harmonized across most regulatory regions in the world is a big advantage in comparison to the narrower spectrum available in other frequency bands, like 2.4 GHz and 5 GHz, available for 802.11 standards. Both ECMA and 15.3c employ a channel plan that consists in

 ISBN: 978-1-61208-203-5

dividing the available spectrum into 2.16 GHz frequency bands for each channel.

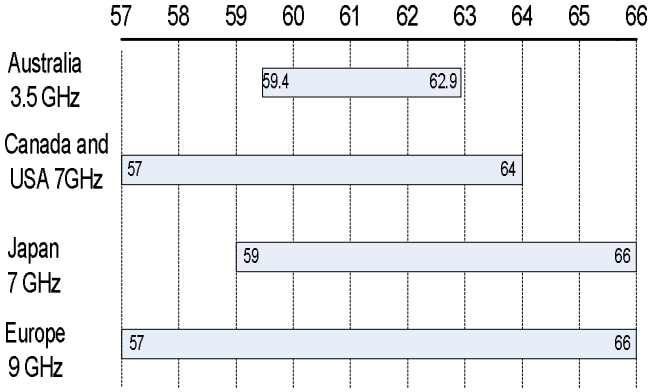


Figure 1. The available spectrum for indoor wireless communications in the 60 GHz band around the world

Both technologies present high perspectives for future implementations, offering important benefits to the users. They both try to enhance the user experience by offering higher data rates, one for wireless local area networks, the other for radio mobile communications.

Given the tendency of developers to integrate and unify the current technologies, the 60 GHz WLAN and femtocell networks, offer real perspectives, but it depends on the users choice, regarding which will have the best advantage.

The rest of the paper is organized as follows: in Section II, we present a comparative study between the two technologies, regarding some technical aspects encountered in the implementation process. In Section III, we consider a case study involving an indoor office environment in which the two networks will be implemented and studied. An analysis is done regarding aspects like the obtained coverage, the resulting interference or the attenuations introduced by the environment objects. Finally, Section IV concludes the paper and presents some future aspects.

II. COMPARISON OF TECHNICAL ASPECTS

The fact that both technologies address the same market segment, i.e., that of the clients situated in the indoor environment which need enhanced data rates for communication, constitutes an important start point in developing a solid comparison. Being in the early stages of network implementations, offers the possibility to have a much wider view concerning the research in these domains. In this way, the development of common points can be realized, leading in the future to the integration of these two types of technologies, in order to enhance the quality of service experienced by the client.

In this section, we will emphasize the main characteristics of the two technologies regarding some key aspects like the integration with the existing implementations, the connectivity to the current networks, mobility and handover possibilities and also interference within the deployed network or with the existing one.

A. *Integration with the existing implementations*

Regarding the 60 GHz WLAN technology, possibilities of integration with existing 2.4 GHz and 5 GHz wireless LANs, represent a major research topic; this is primarily due to the high costs, which result in the implementation of a purely 60 GHz network, as we will see later in the paper. The authors of [9] present a method to integrate the three WLAN technologies, facilitating the commercialization of equipment operating in triple band. Using this, the client can choose the operating frequency depending on the needs and the environment: 2.4 GHz for applications with reduced needs, 5 GHz for applications regarding confidential traffic and 60 GHz for high transfer rate applications. Thus, a mandatory enhancement of the equipment and terminals used must be done, in order to facilitate the implementation of the new 60 GHz technology. This, however, would not be an easy task given the fact that most of the current access points operating in the 2.4 GHz and 5 GHz frequency bands, have omnidirectional antennas, which, in the case of the 60 GHz technology, is not a well suited option because of the high attenuation of the waves transmitted on this frequency. Thus, especially for Non-Line of Sight (NLOS) communications, antennas with higher gains are mandatory in order to reach the receiver. The use of the omnidirectional antennas for the 60 GHz operating frequency would be viable only for direct Line of Sight (LOS) communications, which is not always the case in real life scenarios. The addition of new antennas will rise up the costs, leading to a poorer interest in the technology. Therefore, different approaches need to be found.

In the case of the femtocell networks, the transmitters must integrate into the existing cellular architecture with no modifications to the first. Thus, the existing terminals must be able to connect to the femtocell with no enhancements needed. The fact that the femtocells use the same operational frequency as the macro network is an important advantage. However, architectural modifications need to be done in order to cope with the femtocell concept. Given the opportunistic nature of the femtocell deployment, meaning that the femtocell base stations are implemented by the user, and not by the operator, one may not be able to predict their location; thus, radio planning simulations, prior to the actual deployment, can not be done in order to enhance the operation. Therefore, a new entity must be defined in order to manage and enhance the functionality of femtocells among them, and within the core network. This entity is called the Femto GW and it is the preferred option by the standardization bodies [10]. Its main role is to manage and control the operation of the active femtocells. Among its functionalities are: assuring a secure connection between the femtocell base station and the core network (CN), providing support for paging and handover procedures, transparent transfer of Layer 3 messages between the User Equipment (UE) and core network. A more detailed description of the structure and roles of the Femto GW or Home NodeB GW, in the 3GPP terminology, is given in [11] and [12]. The Femto GW interfaces towards the other entities of the network are defined in [13], [14] and [15].

 ISBN: 978-1-61208-203-5

The dual mode operation Wi-Fi/ 3G is not needed in the case of using femtocells, but this could be a further research topic regarding the integration of femtocells and 60 GHz equipment within the same device.

B. *Connectivity to the current networks*

Maybe the most important issue regarding the 60 GHz WLAN concept is represented by the short coverage area of a transmitter. It is well known that the waves emitted within the 60 GHz frequency band are very much attenuated by the surrounding environment and also by the oxygen, because of the fact that this frequency is the resonance frequency for oxygen. A detailed study of the attenuations involved is presented in [8]. Therefore, a concrete wall, for example, acts as an isolator for the radio waves, introducing an attenuation of up to 40 dB [16], depending on the width of the obstacle. Practically, in order to implement a 60 GHz WLAN network, a transmitter needs to be implemented in each room of the indoor environment. Thus, we can clearly say, without creating an abuse of terms, that the 60 GHz WLAN acts as a cellular WLAN. In [16], Genc et al. present an architecture for this kind of network, in which the transmitters are connected through fiber optics. Considering this, we can establish a common point between the femtocell concept and the 60 GHz WLAN, taking into consideration that the femtocell network is connected to the core network of the operator through a similar backhaul connection. Figure 2 presents a practical generic architecture that can be implemented for both technologies.

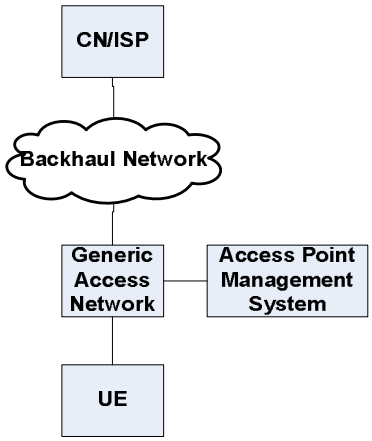


Figure 2. Generic architecture for 60 GHz and femtocell networks

In each case, an AP-MS (Access Point Management System) must be developed in order to coordinate the functioning of the transmitters. Both types of technologies use a backhaul connection in order to connect to the existing infrastructure, i.e., the core network (CN) in the case of the cellular communication network, and respectively the network of the Internet Service Provider (ISP) for the case of the wireless local area network. The existing mechanisms to integrate an AP into a WLAN must be modified in order to cope with the characteristics of the 60 GHz technology. The access mechanisms used until now in WLANs, CSMA/CA can no longer be used given the fact that due to the isolation created by the environment, the 60 GHz cells will have very little superimposing of the coverage areas, thus the receivers can detect only one transmitter in any point in the environment. In the case of the femtocells, a Femto GW device has been developed which enables the access points to communicate with the core network and between them. Practically, a femtocell is seen by the terminal as another macrocell, and therefore the handover process may remain the same, except for the fact that it is realized through the Femto GW.

C. *Mobility and handover possibilities*

When a wireless local area network is implemented in a specific location, the user must have the desired mobility rights. Given the fact that the superimposing of the coverage areas is very small in the 60 GHz technology, especially near windows or doors, the handover process from one transmitter to another needs to be done in this area. Thus, a very fast handover procedure needs to be done in order to maintain the connection of a user passing from one room to another. One such solution is given in [16], which proposes that WLAN cells should be grouped in such a way that all the cells in a specific group transmit the same information on the same channel. Using this, we obtain larger cells that may be planned easier, in such a way that the superimposing of the coverage areas is enlarged, making it easier for the handover procedure to be realized.

In the case of the femtocell concept, things are different. It mainly depends on what access mechanism the femtocells use: in open access mode, all the users of the macrocellular network are able to connect to the femtocell, thus a handover procedure can be done; in closed access mode, the outside users are not allowed to connect to the femtocell device, and in this case the FAP acts as an important interference source. The scientific literature proposes also a hybrid mode, in which full access is given to the registered users (subscribers), while the non-subscribers are allowed only limited access to the resources, for minimal applications [17]. However, even when considering the open access mode, the large number of handovers which a mobile user may experience while passing through the coverage areas of several femtocells, may lead to increased signaling on the network, which degrades the performance. The authors of [6] present an algorithm which may be used in order minimize the core network signaling.

D. *Interference Issues*

One important issue in designing any cellular network is represented by the interference which occurs between the transmitters, at the receivers site.

In the case of the purely 60 GHz WLAN, interference is not a problem given the fact that a cell created by a transmitter is isolated by the environment obstacles. This is due to the high attenuations created by the objects in the surrounding environment on the waves operating on this frequency. In the case of a combined 2.4 GHz and 60 GHz network, the principle is the same for the 60 GHz transmitters, while for the ones operating at 2.4 GHz, the

 ISBN: 978-1-61208-203-5

access mechanism used, CSMA/CA avoids the negative impact of the interference between the transmitters.

However, this issue is of critical importance while implementing a femto-macro network. Here, interference between the femtocell and macrocell layers occurs, due to the closed access mechanism implemented in the femtocell. Using this, only subscribers are allowed to connect to the femtocell. For the other macrocell users that enter the coverage area of a femtocell, this acts as a powerful interference source which can degrade de QoS experienced by the user in such a way that it could lead to outage. Other types of interference are represented by the interference caused by a MacroBS to user that is connected to a FAP and results in a downgrade of the SINR level; interference caused by a macro user on the uplink communication, to a FAP, or even femto-to-femto interference which can occur in dense urban areas where femtocells can be deployed close to each other. In the femtocell domain, ways of reducing the cross- and co-tier interference represents the major research topic. Several possibilities have been presented in papers like [18]-[21]. However, a stable and final solution has not yet been found, but research is currently making important progress.

III. CASE STUDY DEPLOYMENT

In this section, we will concentrate on the behaviour of the two technologies described above, from the radio propagation point of view. Therefore, we will analyze the impact of deploying both the 60 GHz WLAN access points and the cellular communications femtocell access point. In order to have a better understanding of the impact resulted from the deployment of each technology, we will consider the same environment conditions for both cases.

The scenario involved in this experiment consists of an indoor office environment, in which the two technologies will be deployed. The environment and the simulations are realized using the RPS (Radiowave Propagation Simulator) program [22], a tool which is no longer available under this brand, but the same functionalities are encountered in the tool provided by Actix [23].

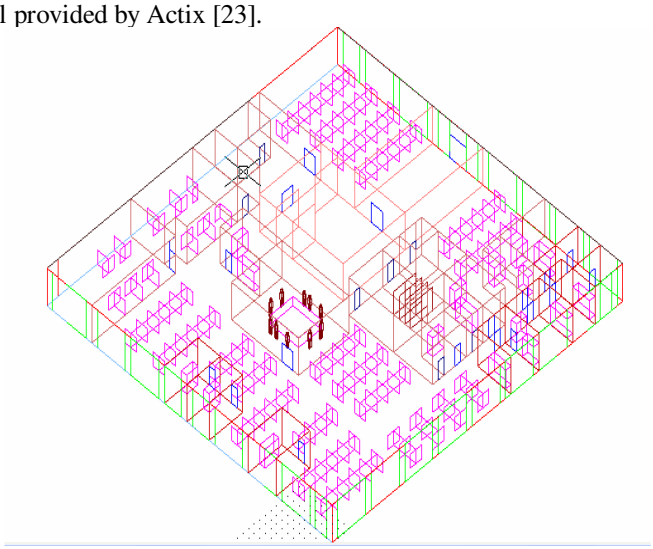


Figure 3. The deployment scenario

The medium is built in the Environment Editor, a program similar to AutoCAD [24], which enables the construction to be realized on layers, each of them being characterized by a series of parameters: thickness, electrical permittivity, the possibility to allow or not penetrations, diffractions or reflections. The environment consists of an office scenario in which the effects of the presence of the outer and inner walls, of the windows and doors, are taken into consideration in the evaluation of the coverage. The environment is very complex from the point of view of the types of materials used: concrete, brick, reinforced wood, wood, glass, and we have even simulated the presence of the human body, which has an important effect especially on the 60 GHz wireless local area network. The created scenario is illustrated in Figure 3.

In order assure a sufficient coverage of the environment using only transmitters that must operate at the 60 GHz frequency, the resulting network would be extremely costly, resulting in a number of up to 27 transmitters. The technical parameters of each one are presented in Table 1, taken from [4].

TABLE I. TECHNICAL PARAMETERS OF THE 60 GHZ TRAMSMITTERS

Tx name	Antenna type	Orientation of the antenna		Antenna height [m]	Tx power [dBm]
		Phi	*Theta*		
Tx1	Omni	0	0	3	12
Tx2	Omni	0	0	3	12
Tx3	Patch	45	0	3	12
Tx4	Patch	330	0	3	12
Tx5	Patch	280	0	3	12
Tx6	Omni	0	0	3	12
Tx7	Patch	135	0	3	12
Tx8	Omni	0	0	3	12
Tx9	Omni	0	0	3	12
Tx10	Patch	220	0	3	12
Tx11	Omni	0	0	3	12
Tx12	Patch	45	0	3	12
Tx13	Omni	0	0	3	12
Tx14	Omni	0	0	3	12
Tx15	Patch	270	0	3	12
Tx16	Patch	270	0	3	12
Tx17	Patch	110	0	3	12
Tx18	Omni	0	0	3	12
Tx19	Omni	0	0	3	12
Tx20	Omni	0	0	3	12
Tx21	Omni	0	0	3	12
Tx22	Omni	0	0	3	12
Tx23	Patch	270	0	3	12
Tx24	Omni	0	0	3	12
Tx25	Horn	260	0	3	12
Tx26	Patch	180	0	2	12
Tx27	Horn	90	0	2	12

A sample snapshot of the level of the received signal strength for the deployment of the 60 GHz WLAN access points is presented in Figure 4. One may notice the strong attenuations introduced by the environment upon the signal waves transmitted with the 60 GHz frequency. Practically, in order to assure the coverage in all the indoor environment, at least one transmitter is necessary in each closed area.

 ISBN: 978-1-61208-203-5

Moreover the presence of the human body significantly attenuates the level of the received signal strength.

That is why such a solution is not feasible, and the proposed solution presented also in [4], would be to combine the 60 GHz technology with the existing 2.4 GHz network. The placement of the 60 GHz transmitters in only a few points of the environment would ease the implementation and help reduce costs.

However, such a solution needs to be supported by the ability to integrate these two types of wireless LANs. Practically, we have implemented a 2.4 GHz network in all the environment, able to offer services for users that need support for common applications, while the 60 GHz network is implemented in only a few key points of the environment, like conference rooms, executives offices, etc., able to support the need for high data rates applications like video conferences or the transfer of large files.

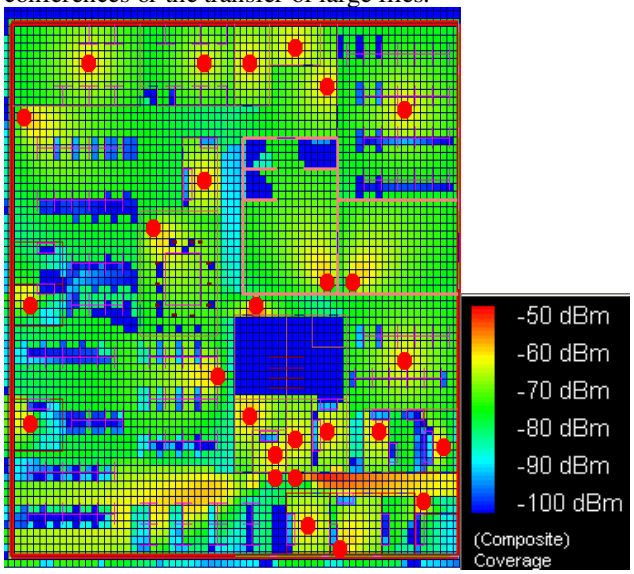


Figure 4. Level of received signal strength for the 60 GHZ WLAN.

The same scenario considering a femtocell-macrocell network would imply a different approach. Considering a macrocell network already deployed in the exterior of the building, the complicated structure of the scenario and simulations done, which reveal a poor indoor coverage of the macrocellular approach, demonstrate the need to implement a femtocell transmitter. By doing this, with only one femtocell transmitter, the coverage inside the office building is assured with good results. The technical parameters of the deployed transmitters are presented in Table 2.

Considering the cellular network implementation standard as being UMTS, the operating frequency chosen is considered to be 2 GHz. The considered macrocellular base station is placed at a distance of 550 meters from the indoor environment, while the femtocells are placed in the environment, at various positions such that they will assure a sufficiently high level of the received signal strength . The considered transmit power is 43 dBm for the MacroBS and 20 dBm for the FAPs. In case of the FAPs, an adaptive power control algorithm would be necessary to reduce the cross-tier interference that occurs between the femtocells and macrocell respectively.

TABLE II. TECHNICAL PARAMETERS OF THE CELLULAR NETWORK TRANSMITTERS

Parameter	MacroBS	FAP
Antenna type	UMTS 30.03 Sector antenna	Omnidirectional
Antenna Gain	1 dB	0 dB
Polarization	Linear vertical	Linear vertical
Orientation of the antenna	Phi = 90 deg.	Phi = 0 deg.
	Theta = 0 deg	Theta = 0 deg
Antenna height	7 meters	3 meters
Transmit power	43 dBm	20 dBm
Carrier frequency	2 GHz	2 GHz
Distance to the indoor location boundary (window)	550 meters	Variable, depending on position

A sample snapshot of the level of the received signal strength coming from the femtocell base stations is presented in Figure 5.

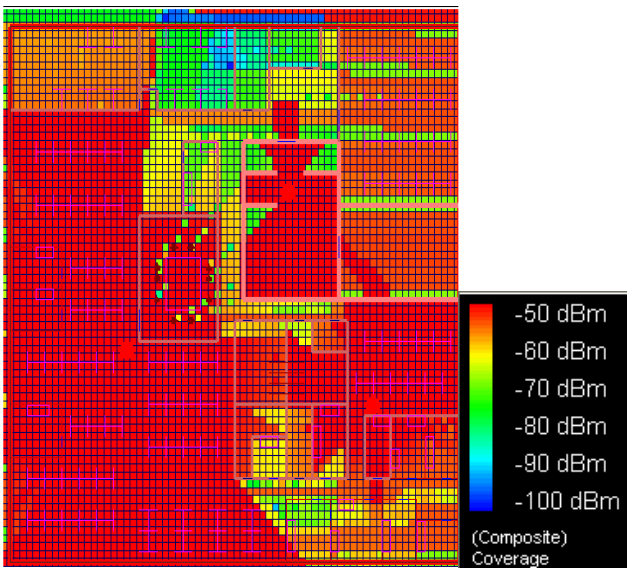


Figure 5. Level of received signal strength for the femto-macro network.

One may notice that in order to assure coverage inside the indoor environment only three femtocell base stations are necessary, considering also that we benefit from the outdoor signal of the macrocell base station. Therefore, from the point of view of the user, the cost are significantly lower in the case of using the femtocell solution, rather than the 60 GHz WLAN, mainly because of the much lower number of transmitters needed to cover that certain area.

The complex nature of the environment influences the coverage differently in the cases of the two technologies. Therefore, one important factor because of which we need such a high number of transmitters in case of the WLAN, necessary to cover the scenario, is represented by the attenuation created by the environment to the traveling waves. Table 3, presents a comparative study between the attenuations that occur for the 60 GHz and 2 GHz,

 ISBN: 978-1-61208-203-5

respectively. One interesting fact here, consists in the attenuation introduced by the human body to the waves operating at 60 GHz. As mentioned before, the 60 GHz frequency is the resonance frequency for oxygen, and given the fact that the human body is constituted in a high proportion out of water, such an obstacle practically creates isolation to a receiver situated behind it.

TABLE III. ATTENUATIONS INTRODUCED BY THE ENVIRONMENT

Obstacle	Attenuation introduced [dBm]	
	2 GHz	*60 GHz*
Outer wall (Concrete 40 cm)	27 ~ 30	No detectable signal
Inner wall (Glass 3cm)	3 ~ 4	10 ~ 12
Inner wall (Brick 10 cm)	16 ~ 19	No detectable signal
Door (Glass 2 cm)	2.5 ~ 4	3 ~ 5
Cubicles (Wood 5 cm)	3.5~ 5	18 ~ 23
Door (Reinforced wood 3 cm)	23 ~ 25	No detectable signal
Human body	13	No detectable signal

Considering these results, one important factor when choosing between one technology or the other is represented by the costs of the deployment at the user. Therefore, when implementing a combined 60 GHz – 2.4 GHz WLAN, the user would need to support the entire cost of the equipment. Even though the 60 GHz transmitters are positioned only in a few points in the environment, and with the research done in using the CMOS technology to develop these transmitters, the overall cost of the WLAN network would exceed that of choosing the femtocell technology. In this latter case, the user would need to acquire only the FAP, which is by definition of low cost; thus, the investment is minimal.

With the development of the new cellular standards like LTE and WiMAX combined with the femtocell implementation which assures the necessary radio coverage in the indoor environment, the user would be able to obtain comparable or even higher data rates, than by using the 2.4 GHz WLAN system. Another advantage in favor of the femtocell is that the handsets need no additional improvements in order to work using the femtocell technology considering the same operating frequency, while for the 60 GHz technology the terminals would need additional improvements in order to facilitate this operation.

But, the femtocell technology will never be able to achieve the high data rates offered by the 60 GHz WLAN. Therefore, when choosing one technology or the other the user must decide if it is worth to invest in a costly network, but which offers great transfer rates, or if its requirements can be supported by a less costly network, capable of assuring sufficient transfer rates.

IV. CONCLUSIONS

The goal of the paper was to realize a comparison between two upcoming new technologies that will be available on the market in the next few years: the 60 GHz technology with direct applications in the wireless local area networks domain, and the femtocell technology which will be implemented in order to enhance the coverage and capacity of the existing macrocellular networks.

Both technologies offer good advantages for the users: the 60 GHz WLAN offers great transfer rates, unmatched by any of the existing technologies; while the femtocell concept enhances the coverage of cellular networks leading to a higher QoS level at the receiver site, offering the possibility to obtain good transfer rates, enhances the capacity of the network by managing a part of the users that were normally handled by the macrocell, all of these with the advantage of mobility. Therefore, the femtocell is advantageous for both the operator and the user.

But, besides these benefits, the mentioned technologies have some drawbacks as well: in the case of the 60 GHz network, the major issue refers to the fact that the coverage of a transmitter is limited by the closed environment it is placed in. Also, another relevant problem is represented by the handover of a user between two transmitters, mainly because of the little superimposing of the coverage areas of two adjacent cells..

For the femtocell concept, the major problem relates to the interferences that occur between the femtocellular and macrocellular layers. This issue will probably be resolved in the near future due to the extensive research done in this domain in the last few years.

Therefore, in the mass market implementation the femtocell concept will outrank the 60 GHz WLAN, due to its low cost and mobility advantage that it provides. This does not mean that the 60 GHz WLAN will disappear from the market, on the contrary, its implementation will address more high demanding applications most likely for the technological and research domains.

ACKNOWLEDGMENT

This paper was supported by the project "Doctoral studies in engineering sciences for developing the knowledge based society-SIDOC" contract no. POSDRU/88/1.5/S/ 60078, project co-funded from European Social Fund through Sectorial Operational Program Human Resources 2007-2013.

REFERENCES

[1] http://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/, [retrieved March 2012].

[2] http://standards.ieee.org/about/get/802/802.16.html, [retrieved March 2012].

[3] V. Chandrasekhar, J. G. Andrews and A. Gatherer, "Femtocell Networks: A Survey", The University of Texas at Austin Texas Instruments, June 28, 2008.

[4] C. Androne, T. Palade and E. Puschita, "Open Loop Sensor based System used for Mitigation of Cross-tier Interference in Femtocell Networks ", Proc. of TELFOR 2010, Belgrade, Serbia.

[5] 3GPP TR 36.814, "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects" Version 9, 2010.

[6] J. Zhang and G. de la Roche, "Femtocells, Technologies and Deployments", Wiley, 2010.

[7] http://www.mmwaves.com/products.cfm/product/20-194-0.htm, [retrieved March 2012].

[8] L. L. Yang, "60 GHz : Opportunity for Gigabit WPAN and WLAN Convergence", , Intel Corporation, January, 2009.

 ISBN: 978-1-61208-203-5

[9] C. Androne and T. Palade, "Radio Coverage and Performance Analysis for Local Area Networks', Proc. of ISETC 2010, Timisoara, Romania.

[10] 3GPP TR R3.020, "Home (e)NodeB; Network Aspects," Sept. 2008.

[11] TS 43.318, "Generic Access Network (GAN) Stage 2", Rel-5.

[12] TS 44.318, "Generic Access Network (GAN); Mobile GAN Interface Layer 3 Specification", Rel-5.

[13] TS 25.410, "UTRAN Iu Interface: General Aspects and Principles", Rel-5.

[14] TS 25.450, "UTRAN IuPC Interface General Aspects and Principles", Rel-5.

[15] TS 25.419, "UTRAN IuBC Interface: Service Area Broadcast Protocol (SABP)", Rel-5.

[16] Z. Genc, B. L. Dang, J. Wang and I. Niemegeers "Home networking at 60 GHz: Challenges and Research issues", IFIP International Federation for Information Processing, 2008, Volume 256/2008, pp. 51-68.,.

[17] H. Claussen, L. T. W. Ho, and L. G. Samuel, "Self-Optimization of Coverage for Femtocell Deployments," Proc. Wireless Telecommun. Symposium (WTS '08) (Pomona, CA, 2008), pp. 278–285.

[18] V. Chandrasekhar, J. G. Andrews, T. Muharemovic, Z. Shen and A. Gatherer, "Power Control in Two-tier Femtocell Networks," IEEE Trans. Wireless Comm., vol. 8, no. 7, July 2009.

[19] Z. Shi, M. C. Reed and M. Zhao, "On Uplink Interference Scenarios in Two-Tier Macro and Femto Co-existing UMTS Networks," EURASIP Journal on Wireless Communicaion and Networking, Vol. 2010, Article No. 4, January 2010.

[20] X. Li, L. Qian and D. Kataria, "Downlink Power Control in Co-Channel Macrocell Femtocell overlay," Proc. of 43rd Annual Conference on Information Sciences and Systems, pp. 383 – 388, 2009.

[21] V. Chandrasekhar, M. Kountouris and J. G. Andrews, "Coverage in Multi-Antenna Two-Tier Networks," IEEE Trans. Wireless Comm., vol. 8, no. 10, 2009.

[22] http://www.mindservices.com.au/index.php?option=com_content&view=article&id=85:rf-planningtoolls&catid=65:library-references &Itemid=61, [retrieved March 2012].

[23] http://www.actix.com/our-products/radioplan/index.html, [retrieved March 2012].

[24] RPS manual, available with RadioWave Propagation Simulator kit.

 ISBN: 978-1-61208-203-5

A Greedy-based Network Planning Algorithm for Heterogeneous Smart Grid Infrastructures

Christian Müller
TU Dortmund University
Communication Networks Institute (CNI)
44221 Dortmund, Germany
Email: christian5.mueller@tu-dortmund.de

Christian Wietfeld
TU Dortmund University
Communication Networks Institute (CNI)
44221 Dortmund, Germany
Email: christian.wietfeld@tu-dortmund.de

***Abstract*—Focus of this paper is the evaluation and optimization of automatic network planning algorithms considering different communication technologies supporting Smart Grid communication infrastructures. Therefore, a performance evaluation and sensitivity analysis of parameters of greedy-based algorithms solving the covering-location problem are implemented and analyzed in a discrete-event simulation environment. Based upon the presented results, an optimization based on the greedy algorithm is introduced considering Smart Grid technology and topology specific parameters. An evaluation for several real-world reference scenarios shows the influence of multi-layered and heterogeneous network topologies, which are typically used in Smart Gird ICT networks, including wired, wireless and Powerline Communication technologies. Depending on the technology, an optimization of the deployment level and number of network entities can be achieved and is presented by a reduction up to 30% for single-technology topologies and up to 10% for heterogeneous topologies.**

Keywords-network planning algorithm; covering location problem; heterogeneous infrastructures; smart grid.

I. Introduction

Current Smart Grid approaches comprise an actively integration of distributed energy sources and loads into the energy grid in order to enable a more balanced usage of volatile energy sources and movable load systems. In this context, several smart energy management approaches are present like locally managed and self-sustaining Micro Grids [1] and centralized load coordination like Demand Side Management (DSM), Distributed Energy Resources (DER) for example based on dynamic energy prices. At this point, seamless integration of DER and DSM at the customers households are some of the key capabilities of future Smart Grid infrastructures. For this purpose, the underlying communication infrastructure requires a reliable, sufficient dimensioned and demand-oriented network design in order to transport metering data, control information and to provide added-value services with the required Quality-of-Service (QoS). But the challenging task in designing a network infrastructure for these application is caused by the heterogeneity of access technologies (e.g., GPRS, PLC, DSL), combining shared and dedicated infrastructures, integrating existing networks, deploying new networks and providing Home Area Networks (HAN), Neighborhood Area Networks (NAN) and Wide Area Networks (WAN) for different application scenarios [2]. The variety in applicable technologies leads to a heterogeneous infrastructure, which requires detailed and adjustable network planning algorithms considering the different architectural structures for different technologies.

The work in this paper concentrates on the design and evaluation of heterogeneous network infrastructure planning algorithms. Therefore several network planning methologies are discussed in terms of cell size, cell capacity, multi-layered and heterogeneous topologies. A Greedy-based algorithm is evaluated by simulation and enhanced for typical Smart Grid infrastructures.

The paper is structured as follows: Section II introduces related work in terms of ongoing Smart Grid projects and network planning algorithms. The implemented algorithms and the simulation environment are presented in Section III. The performance evaluation of a typical Smart Grid scenario is addressed in Section IV. Finally, the paper is finished with conclusion and an outlook on future work.

II. Problem Statement and Related Work

An overview on heterogeneous network topologies supporting the Smart Gird by using different technologies for multiple application scenarios is given in the following Section II-A. An approach describing the optimization problem is given in Section II-B, whereas a state-of-the-art review of network planning algorithms is provided in Section II-C.

A. Smart Grid Topologies

Network infrastructures for Smart Grids applications comprise different aggregation levels in HAN, NAN and WAN infrastructures in order to support different ICT and Energy components [3]. Figure 1 shows a multi-layered network topology for integrating large-scale components, like wind

 ISBN: 978-1-61208-203-5

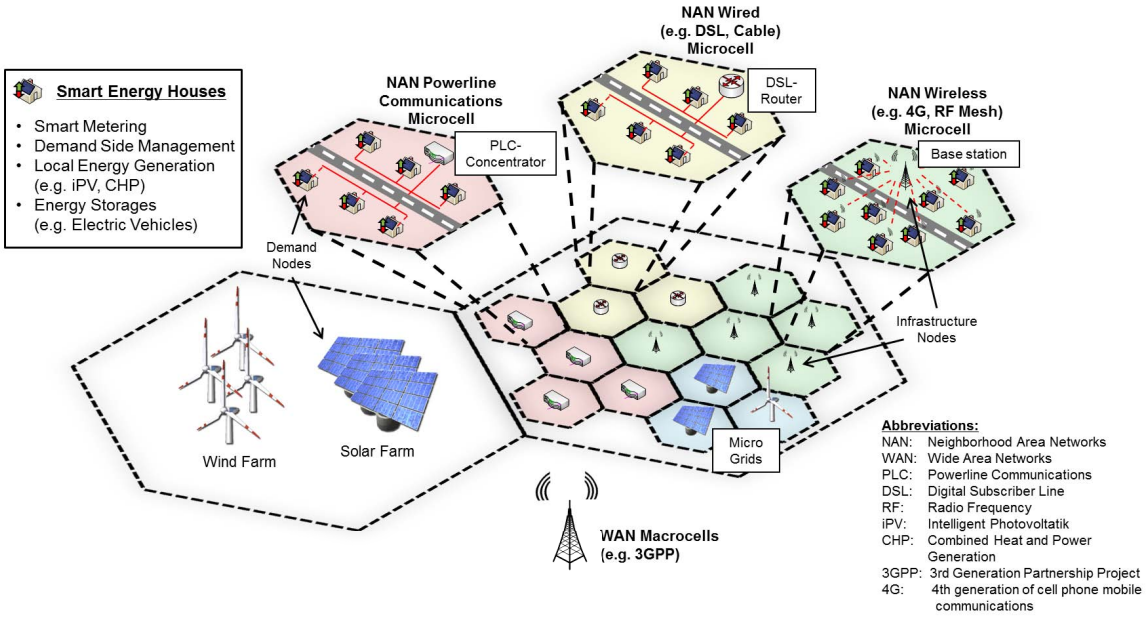


Figure 1. Multi-Layered Network Topologies for Smart Grid Application Scenarios

and solar farms, as well as particular smart energy households and micro grids. In order to provide connectivity to the customers premises equipment, several options has been introduced [4], which mainly base on two scenarios: a *dedicated* network infrastructure and a *shared* network infrastructure. By using a shared network infrastructure, restrictions in terms of Quality-of-Service have to be accepted due to non-exclusive usage of the medium. On the other hand, this solution offers an economic alternative by using existing infrastructures.

Wired preexisting technologies (e.g., DSL, FTTx, GPON) [5] in the Smart Grid context are used for integrating the prosumers (**pro**ducer and co**sumer**) households and sub-, resp. transformer stations into the infrastructure. Since installation costs are higher compared to wireless technologies, the integration of existing infrastructures is widely preferred.

Using the powerline as transmission medium for data has been discussed decades-long and has been established successfully for the transport grid. Concerning the present efforts on integrating new actors like the prosumers, into the Smart Grid, especially Broadband PLC technologies (IEEE P1901, HomePlug 1.0/1.0 Turbo/AV/AV+, DS2, Panasonic HD) become more relevant. Due to new capabilities like larger bandwidth, higher modulation schemes and notching filter, the BPLC technologies offers an economic solution for the communication infrastructure on several levels of the Smart Grid. Furthermore, narrowband PLC technologies are discussed as a last-mile solution due to moderate installation costs and exclusive usage.

Wireless Technologies, like GSM, UMTS as well as LTE and Mobile WiMAX offer a cost-efficient solution for new communication infrastructures due to the saved installation costs for cables. On the other side, wireless technologies are based on a substantiated network design covering resource and frequency allocation. Several Smart Metering projects are based upon wireless low data rate approaches, but for a comprehensive installation of Smart Meters and for offering enhanced services like DSM enhancements, the usage of next generation cellular networks for machine-to-machine (M2M) services is currently evaluated. Especially the usage of lower frequency ranges (e.g., digital dividend after digital television transition) for dedicated services (Smart Metering, DSM, Substation Automation) offers a promising solution for covering rural areas, whereas the development of future network deployments needs to be taken into account.

B. Covering-Location-Problem

Deploying and optimizing the previously described networking technologies is related to the *Set-Covering-Location-Problem (SCLP)*, which is one of the most studied NP-hard problems [6]. The goal is to cover a given set of demand nodes $J = \{0, ..., m\}$, e.g., Smart Energy Houses, with the minimum number of required infrastructure nodes, e.g., base stations, which are taken from a set of possible infrastructure node positions $I = \{0, ..., n\}$. A mathematical description of the optimization problem is given by

$$min \sum_{i \in I} y_i \quad (1)$$

under the condition that

$$\sum_{i \in I} r_{ij} y_i \geq 1 \quad for \quad j \in J \tag{2}$$

with

$$y_i, r_{ij} \in \{0,1\} \quad for \quad i \in I, j \in J \tag{3}$$

whereas y_i is representing the deployed infrastructure node positions and r_{ij} is indicating the according covered demand nodes in range.

The *Maximum-Covering-Location-Problem (MCLP)* is a modification of the SCLP, where the goal is to cover a maximum number of demand nodes with a limited number of infrastructure nodes. In order to priorize the demand nodes an additional parameter b_j for $j \in J$ is introduced and the relation for the optimization problem is given by

$$max \sum_{j \in J} b_j \cdot x_j \tag{4}$$

under the condition that

$$\sum_{i \in I} r_{ij} \cdot y_i \geq x_j \quad for \quad j \in J \tag{5}$$

and

$$\sum_{i \in I} y_i = p \tag{6}$$

and

$$x_j, y_i \in \{0,1\} \quad for \quad i \in I, j \in J \tag{7}$$

Several approaches for solving the optimization problem are discussed in literature and summarized in the next section.

C. Network Planning Algorithms

In general, the following networking algorithms are discussed in the literature [7]: *Exact Algorithms*, *Genetic Algorithms* and *Heuristic Algorithms*.

In order to obtain an optimal solution for the problem, exact algorithms search all potential solutions in the parameter space, which is a time-consuming procedure, in the context of Smart Grid infrastructures, where thousands of nodes are taken into account [8]. Another approach are genetic algorithms based on Darwin theory of natural selection and its class of evolutionary algorithm. The idea behind this algorithm is to start with an initial population and then individuals from the initial population are selected in order to generate new individual solutions [9][10]. Heuristics algorithms are approximated algorithms and provide relatively *optimal* solutions in a reasonable computation time. This is a compromise between solution quality and execution time, which offers a sufficient solution for network planning. There are several approximated algorithms used to solve network planning problems, like Tabu Search [11], Simulated Annealing [12] and Greedy Algorithm [7], which are iterative algorithms calculating stepwise the local optimum.

III. Simulation Environment

In order to compare the different optimization approaches and network planning methods, a geo-based simulation environment [13] is used for the conducted performance analysis. Due to the geo-based positions of the communication nodes, close to real world scenarios were investigated in order to analyze the impact of the algorithms on the real-world scenarios.

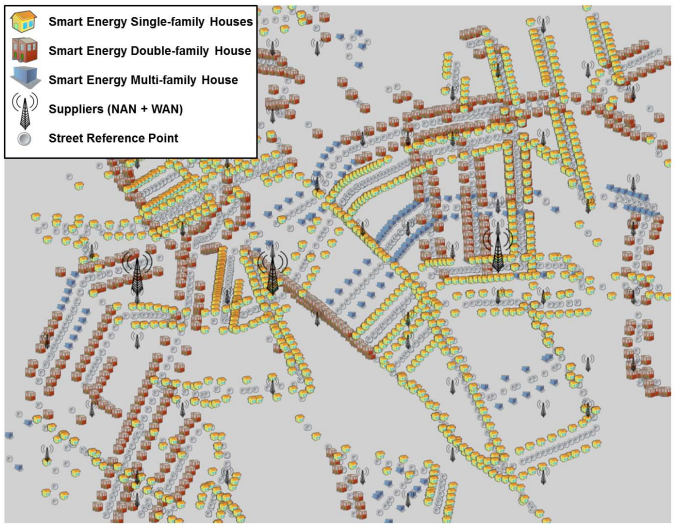


Figure 2. Simulation Scenario with Multi-Layered Infrastructure and Adaptive Cell Size

A. Geo based Scenario Generator

The presented simulation model is based on the discrete event simulator OMNeT++ [14]. The developed geo-position scenario generator acquires the coordinates from different offline sources. This includes own acquired data and governmental/commercial data products, as well as online sources, like the Google Maps API (Premier) or OpenStreetMap. Due to the limitations in requesting data from the online sources, the online procedure is reasonable for smaller scenarios, but in case of a large-scale scenario, the acquisition of the geo-positions can be performed by offline sources. In order to generate a large-scale network with thousands of nodes, a dynamic network creation is necessary. This avoids a

	Broadband NAN Technology		
	Wired	Wireless	PLC
max. Demand Nodes per Cell	96	200	48
min. Demand Nodes per Cell	18	6	6
max. Distance/Range	500 m	200 m	100 m
Infrastructure Node Positions	Streets	Houses	Streets
Coverage Area Calculation	Distance	Channel Model	Distance

Table I
Parameters for different application scenarios

 ISBN: 978-1-61208-203-5

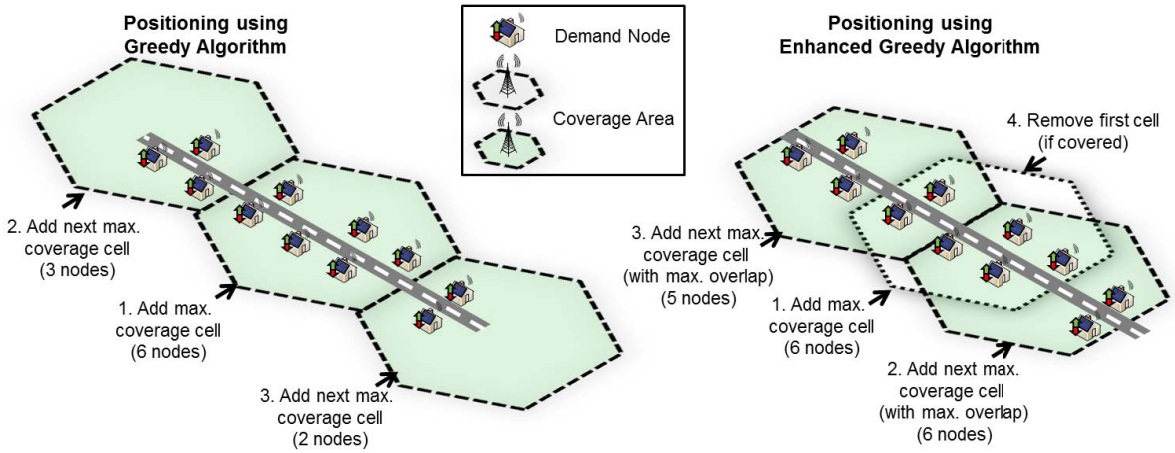


Figure 3. Network Planning based upon Greedy Algorithms and Enhanced Greedy Algorithms

manual, time-consuming static generation and configuration of each node. In our approach, we use geographic positions of real locations, e.g., houses, as input parameters for the automatic network generation. This ensures a close to reality network topology. An exemplary geo-based simulation scenario is presented in Figure 2.

Initialization, generation and configuration of the network are executed by a core simulation [13]. A set of preconfigured geographic positions (north-west and south-east corner coordinates) mark the simulation playground. The core simulation can receive the positions of nodes, which are located within the playground, from an offline source and online from an external SQL database, e.g., the geo-position database *GeoDatabase*. Via a connection between the core simulation and the GeoDatabase [13], all information about existing nodes within the playground and neighbors of particular nodes can be retrieved, including the information listed in Table III-A. The dynamic node creation is based on the received geographic positions. Nodes are placed on the Cartesian positions (x,y), which are calculated using Mercator projection of GPS position data of real locations.

B. Network Planning Algorithms

The network planning algorithm is based upon a *Greedy-Adding Algorithm* [15] with several adjustments in terms of cell size, link budget and multi-layered infrastructures in order to meet the requirements on a communication infrastructure for a Smart Grid. The parameters of implemented different broadband NAN technologies are summarized in Table III-A.

1) Greedy Adding Algorithmic: The Greedy Adding Algorithm offers a powerful approach for solving the CLP, but in some cases a non optimal result is calculated. Especially in areas of high density with a linear placement of communication nodes, which is usually the case for streets or smaller villages, an optimization of the Greedy-Adding algorithm can reduce the overall number of communication cells. Figure 3 shows the comparison of the common and enhanced Greedy Algorithm. The common Greedy Algorithm selects the next best candidate position for an infrastructure node by selecting the position, which covers the maximum number of uncovered demand nodes. This procedure comes along with two disadvantages: On the one hand, two or more equal positions can not be distinguished due to the next best position is chosen from the list. On the other hand, one position with less uncovered demand nodes could have a better coverage due to already covered demand nodes within the area, which are not taken into consideration. The enhanced algorithm adds in a first step an additional condition to the next best candidate selection by choosing the best position with additionally covering already covered demand nodes. In a second step all candidate positions are checked for multiple covered demand nodes and removed if more than one candidate position is available. Finally, the increased overlapping areas of neighbored cells caused by the additional condition are reduced by optimizing the cell radius and threshold adjustment (see Section III-B2).

2) Adaptive Cell Radius: Usually, the initial connectivity range of a candidate point is adjusted by a simple distance calculation. In reality, more parameters are influencing the connectivity range, e.g., transmission power, outdoor-to-indoor transition, fast fading, incident angle, as well as maximum capacity. In order to meet the requirements for a sufficient simulation of the transmission range per cell, the cell radius is adjusted dynamically by increasing the transmission power from a minimum threshold to a maximum threshold until the maximum transmission power or number of traffic nodes is reached.

 ISBN: 978-1-61208-203-5

3) Link Budget: The calculation of the link budget is accomplished by using appropriate analytic channel models, which take into account the distance, incident angle, house type and orientation. In the presented simulations, analytic radio propagation models are used for different topologies described in Section II-A, which enables a large-scale analysis of the topologies and offers extensibility in terms of technologies. For urban areas, the transmission range is calculated by the Okumura-Hata channel model [16][17][18] covering a high density of stationary communication nodes, prevailing Non-Line-Of-Sight connections and smaller communication ranges. In suburban and rural areas, a predominant number of communication nodes are placed in a (Near)Line-of-Sight conditions and usually the communication ranges are increased.

4) Multi-layered Network Topologies: By introducing multi-layered network topologies, e.g., aggregation of *Smart Metering* data by NAN technologies, an overlapping WAN-technology is required for collecting data from the aggregation points. Additionally, nodes, which are not covered due to their secluded position, can also be covered by the overlapping WAN technology and additionally, an economic threshold of minimum numbers of demand nodes per cell can be defined for the network planning process (e.g. min. 6 nodes). The network planning process for the next higher layer uses the same metric as described in Section III-B2 with according parameters for the WAN technology.

IV. Performance Analysis

The results from the simulation are presented in this section. In Figure 4, the coverage of demand nodes is shown

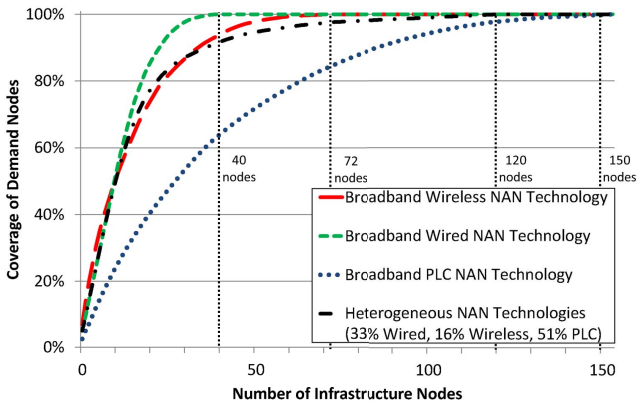


Figure 4. Coverage of Demand Nodes Depending on the Number of Infrastructure Nodes

depending on the number of infrastructure nodes for a single layered topology. Due to the lower capacity of the PLC cells, up to 150 infrastructure nodes are required in order to cover the whole area, whereas the wireless NAN network requires 72 infrastructure nodes and the wired NAN network requires 40 infrastructure nodes. The influence of the transmission range and capacity is shown by the comparison of the wireless and wired NAN technology. Hence, the wireless NAN technology covers more demand nodes during the first planning steps due to the higher capacity, whereas the wired technology shows overall lower number of infrastructure nodes due to the higher transmission range.

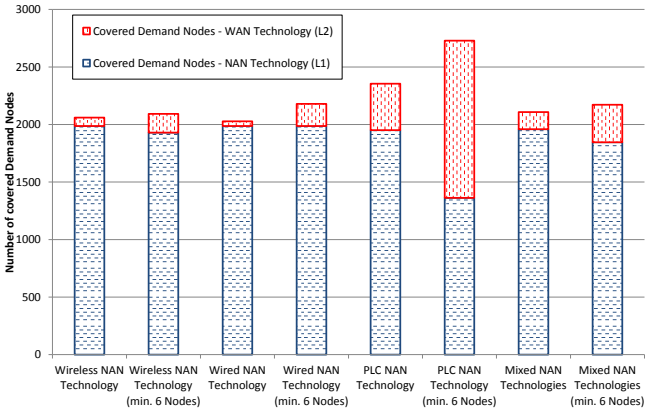


Figure 5. Comparison of NAN Technologies for Multi-Layered Topology Planning

In order to analyze the influence of multiple technologies in one scenario, an additionally heterogeneous network scenario has been analyzed. The results show, that the heterogeneous NAN network shows the same behavior up to a level of deployment of 50% of the wired network, up to a level of deployment of 90% of the wireless network and above 90% of the PLC network. Due to the heterogeneous approach, the number of infrastructure nodes is reduced to 120 together with enhancing the coverage throughout the planning process.

The problem by providing full coverage are secluded nodes (e.g., rural areas). Therefore, a multi-layered infrastructure is introduced by dividing the network into two layers, whereas L1 represents the NAN connectivity and L2 presents the WAN connectivity. The results for a comparison between a single-layered and multi-layered scenarios are shown in Figure 5.

In the multi-layered scenarios, a threshold is defined, which sets the minimum number of demand nodes per NAN cell. All uncovered demand nodes from the L1 are covered in a second planning step (L2) with all L1 infrastructure nodes. The multi-layered planning algorithm shows a better result by reducing the number of infrastructure nodes of up to 30% in the PLC scenario and up to 10% in the mixed scenario.

V. Conclusion and Future Work

The results presented in this paper show the influence of different parameters and enhancements on greedy based net-

 ISBN: 978-1-61208-203-5

work planning algorithms in order to evaluate the usability for Smart Grid ICT topologies. Based upon a real-world scenario, which has been evaluated by a geo-based simulation environment, the influence of multi-layered topologies where analyzed. Hence, a reduction of infrastructure nodes in the presented heterogeneous scenario of up to 10% could be achieved. The influence of a more detailled model of the actual transmission medium by analytic channel models, link budget calculation and adaptive cell size were analyzed as well.

Future work will focus on the integration of more detailed traffic analysis of the particular network entities in order to optimize the maximum cell size an capacity. Furthermore the performance evaluation of the designed networks within a protocol simulation environment and comparison to established telecommunication networks will give some indications of further optimization approaches in terms of adjusting the multi-layer thresholds and scalability issues.

VI. ACKNOWLEDGMENT

The work in this paper was partly funded by the German Federal Ministry of Economics and Technology (BMWi) through the projects E-DeMa (reference number 01ME08019A) [19]. The authors would like to thank the project partners RWE, Miele, Siemens, ProSyst, SWK and ef.Ruhr. The authors wish to acknowledge the efforts of Mr. Henning Böttcher during the project.

REFERENCES

[1] E. Mashhour and S. M. Moghaddas-Tafreshi, "A Review on Operation of Micro Grids and Virtual Power Plants in the Power Markets," in *2nd International Conference on Adaptive Science & Technology (ICAST)*. IEEE, 2009, pp. 273–277.

[2] IEEE P2030/D2.1 Draft, *Draft Guide for Smart Grid Interoperability of Energy Technology and Information Technology Operation with the Electric Power System (EPS), and End-Use Applications and Loads*. New York, USA: IEEE, May 2010.

[3] C. Wietfeld, H. Georg, S. Gröning, C. Lewandowski, C. Müller, and J. Schmutzler, "Wireless M2M Communication Networks for Smart Grid Applications," in *European Wireless 2011 (EW)*, Vienna, Austria, April 2011.

[4] C. Wietfeld, C. Müller, J. Schmutzler, S. Fries, A. Heidenreich, and H.-J. Hof, "ICT Reference Architecture Design based on Requirements for Future Energy Marketplaces," in *1st IEEE International Conference on Smart Grid Communications (SmartGridComm)*. Gaithersburg, Maryland, USA: IEEE, October 2010, pp. 315–320.

[5] C. H. Hauser, D. Bakken, and A. Bose, "A Failure to Communicate: Next Generation Communication Requirements, Technologies, and Architecture for the Electric Power Grid," *IEEE Power and Energy Magazine*, vol. 3, pp. 47–55, March 2005.

[6] K. Tutschku, *Models and Algorithms for Demand-oriented Planning of Telecommunication Systems*. University of Würzburg, July 1999.

[7] R. M. Whitaker and S. Hurley, "Evolution of Planning for Wireless Communication Systems," in *36th Annual Hawaii International Conference on System Sciences (HICSS)*. Big Island, Hawaii, USA: IEEE, January 2003, pp. 10–19.

[8] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 4th ed. The MIT Press, 2001, vol. 1.

[9] Y. Wu and S. Pierre, "Base Station Positioning in Third Generation Mobile Networks," in *Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 2003.

[10] Y. Choi, K. Kim, and N. Kim, "The Positioning of Base Station in Wireless Communication with Genetic Approach," in *Personal Wireless Communications*. Springer, 2007.

[11] M. St-Hilaire, S. Chamberland, and S. Pierre, "A tabu search algorithm for the global planning problem of third generation mobile networks," *Computers & Electrical Engineering*, vol. 34, no. 6, pp. 470–487, 2008.

[12] J. Dréo, A. Pétrowski, and E. Taillard, *Metaheuristics for Hard Optimization: Methods and Case Studies*, P. Siarry, Ed. Springer Verlag Berlin Heidelberg, 2006, vol. 1.

[13] C. Müller, H. Georg, and C. Wietfeld, "A Modularized and Distributed Simulation Environment for Scalability Analysis of Smart Grid ICT Infrastructures," in *5th International Workshop on OMNeT++ co-located with International ICST Conference on Simulation Tools and Techniques (SIMUTools)*. Desenzano, Italy: ICST, March 2012.

[14] A. Varga and R. Hornig, "An Overview of the OMNeT++ Simulation Environment," in *First International Conference on Simulation Tools and Techniques for Communications, Networks and Systems (SIMUTools)*. Marseille, France: ICST, March 2008, pp. 60:1–60:10.

[15] R. Church and C. ReVelle, "The Maximal Covering Location Problem," *Papers in Regional Science*, vol. 32, no. 1, pp. 101–118, January 1974.

[16] H. Okamoto, K. Kitao, and S. Ichitsubo, "Outdoor-to-Indoor Propagation Loss Prediction in 800-MHz to 8-GHz Band for an Urban Area," *IEEE Transactions on Vehicular Technology*, 2009.

[17] E. Damosso and L. Correia, "Digital Mobile Radio Towards Future Generation Systems, COST 231 Final Report," *European Commission*, 1999.

[18] J. Meinilä, P. Kyösti, T. Jämsä, and L. Hentilä, "WINNER II Channel Models," *Wiley Online Library*, 2009.

[19] BMWi Federal Ministry of Economics and Technology, "Project E-DeMa - Development and Demonstration of Decentralized Integrated Energy Systems on the Way Towards the E-Energy Marketplace of the Future," [retrieved: April, 2012]. [Online]. Available: http://www.e-dema.de/en/

 ISBN: 978-1-61208-203-5

Cramer-Rao Lower Bound on RF Pattern Matching Method with Velocity in LTE System

Dengkun Xiao
Beijing Institute, Huawei, Technologies Co., Ltd.
Beijing, China
xiaodengkun@huawei.com

Jie Cui
Beijing Institute, Huawei, Technologies Co., Ltd.
Beijing, China
cuijie@huawei.com

Jiangbo Zhu
Beijing University of Posts and Telecommunications
Beijing, China
zjb349168311@gmail.com

Xinlong Luo
Beijing University of Posts and Telecommunications
Beijing, China
luoxinlong@gmail.com

***Abstract*−The Radio Frequency Pattern Matching (RFPM) method is an useful solution for UE positioning, and it is not sensitive to the channel states and multipath impact compared with the time-of-arrival schemes. With the help of drive test or other estimation algorithms, the RF pattern database can be constructed efficiently. The positioning accuracy of RFPM is tightly related with the member of measurement elements for pattern matching. In this paper, the UE velocity is adopted in the RFPM positioning to improve the positioning accuracy, and the Cramer-Rao lower bound is used for accuracy evaluation and comparison. From the simulation results the positioning accuracy can be improved remarkably when the UE velocity is considered in addtion to the reference signal received power (RSRP), timing advance (TA) and reference signal time difference (RSTD) in the pattern matching.**

Keywords - Cramer-Rao Lower Bound; RFPM

I. INTRODUCTION

Location estimation has received great deal of attention over the last two decades due to the requirements set forced by the US Federal Communications Commission (FCC) for Enhanced-911 (E-911) safety services [1]. RF Pattern Matching is a positioning method proposed in 3GPP RAN4 meetings. This method can locate the target UE in an area by comparing its RF measurement against a detailed model of the RF environment for that area. The RF parameters measured by the UE could be serving cell identity, reference signal strengths for serving and neighboring cells, timing advance, etc. The RF environment model could be stored in the form of a database of signal strengths vectors indexed by location coordinates of an area, and which could be constructed in advance or real time using a combination of RF propagating modeling and possibly test measurements [2].

In [5], a novel method about finger print positioning based on spatial correlation of collected signal samples and spatial diversity is proposed which can improve positioning accuracy and reduce time consumption of constructing a metropolitan-scale radio map, however only the received signal strength (RSS) have been considered in the algorithms which limit the peak performance. In [6], position estimations got from the so-called Neural Network localization are further processed through a Kalman filtering-based tracking algorithm and thereafter, the processed position is matched to the road according to the map-matching technique applied. However, it cannot work without the existing GPS. In [7], a novel positioning system is proposed, which consist of three sub-systems, and the first sub-system solves the problems related to fingerprint localization and involves neural network as key element of the positioning algorithm. It also ignores the distance and velocity information for improving positioning accuracy.

In this paper, we propose to introduce velocity into the existing RF Pattern Matching schemes for performance improvement. If the velocity estimation can be achieved at UE side with specific measurement errors, it can be adopted for RF Pattern Matching to improve the positioning accuracy using the RF database.

This paper is structured as follows. In Section 2, we will introduce RF Pattern Matching positioning method and diversified measurements with Cramer-Rao lower bound error model. In Section 3, based on the analysis, a RFPM method with velocity is presented, and mathematic expression can be provided as a new positioning error model. Section 4 shows simulation results of proposed method compared with traditional RFPM, and the analysis of simulation results show advantages of proposed method. Finally, conclusions are drawn in Section 5.

II. ERROR MODEL ANALYSIS

The general error model is proposed as shown in the paper [2]. It assumes a model where measurements are non-linearly related to the parameter of interest (location, i.e., coordinate of the UE) and are corrupted by Gaussian Noise. More specifically, let y be a length N vector of measurements as below:

$$y = h(x_{HS}) + n \tag{2.1}$$

where $x = (x_{HS1} \quad x_{HS2})$ is the coordinate of the UE, $h(x_{HS}) = (h_1(x_{HS}) \quad h_2(x_{HS}) \quad \ldots \quad h_N(x_{HS}))$ is the N vector of measurements which are the function of UE' coordinate. $n = (n_1 \quad n_2 \quad \ldots \quad n_N)$ is the corresponding noise with the variance.

In estimation theory and statistics, the Cramer-Rao lower

 ISBN: 978-1-61208-203-5

bound (CRLB) expresses a lower bound on the variance of estimators of a deterministic parameter. A location error function of Cramer-Rao lower bound can be calculated as below:

$$\sigma_{LOC} = \sqrt{trace(\sum_i I_i^{-1})} \quad (2.2)$$

$$I_i = \frac{1}{\sigma_i^2} \frac{\partial h_i(x_{HS})}{\partial x_{HS}}^T \frac{\partial h_i(x_{HS})}{\partial x_{HS}} \quad (2.3)$$

(2.3) is the ith measurement function's fisher information matrix. Error of positioning method can be calculated by accumulating all the measurement's information matrix as indicated in [2].

In mobile communication system, many measurement elements including RSRP (reference signal received power), RSTD (reference signal time difference) and TA (timing advance) can be used for RFPM.

The measurement element of RSRP is the estimated value at UE side which can be derived from transmit power and propagation loss. Hata models can be used to represent the RSRP signature models as below:

$$RSRP = RSRP_{REF} - \alpha \times 10 \times \log_{10}(d / d_{REF}) - a \quad (2.4)$$

where α is the pathloss exponent and d is the distance between UE and serving eNodeB, $RSRP_{REF}$ is the reference RSRP calculated from reference distance d_{REF} and

$$a = 0.5 \times FBR \times (1 - \cos(\theta_{HS} - \theta_{CELL}))$$

where FBR is the front-to-back ratio, θ_{HS} is the angle of UE measured positive counter-clockwise from the X-axis, θ_{CELL} is the angle of the antenna boresight measured positive counter-clockwise from the X-axis. So differential coefficient of RSRP can be get as below:

$$\frac{\partial RSRP}{\partial x_{HS}} = \begin{pmatrix} pa_1 - qa_2 & pa_2 + qa_1 \end{pmatrix} \quad (2.5)$$

Where

$$a_1 = \frac{x_{HS1} - x_{CELL1}}{\sqrt{(x_{HS1} - x_{CELL1})^2 + (x_{HS2} - x_{CELL2})^2}}$$

$$a_2 = \frac{x_{HS2} - x_{CELL2}}{\sqrt{(x_{HS1} - x_{CELL1})^2 + (x_{HS2} - x_{CELL2})^2}}$$

$$p = -\frac{10}{In10} * \frac{\alpha}{d}$$

$$q = -0.5 * FBR * \sin(\theta_{HS} - \theta_{CELL}) * \frac{1}{d}$$

(x_{CELL1}, x_{CELL2}) is the coordinate of the eNB of serving cell. RSRP's fisher information matrix can be calculated by using (2.3) and (2.5) as below [3]:

$$I_{RSRP=} \frac{1}{\sigma_{RSRP}^2} \begin{pmatrix} (pa_1 - qa_2)^2 & (pa_2 + qa_1)(pa_1 - qa_2) \\ (pa_2 + qa_1)(pa_1 - qa_2) & (pa_2 + qa_1)^2 \end{pmatrix} \quad (2.6)$$

σ_{RSRP} is the RMS of RSRP measurement error.

RSTD is time difference between positioning signal arrival timing of reference cell and of neighboring cell, which means that it is also the function of coordinate of the target UE. According to RSTD definition, it can be expressed as below [3]:

$$RSTD_i = (d_i - d_1) / c \quad (2.7)$$

where d_i is distance between the UE and *i*-th eNodeB and c is light speed.

$$I_{RSTD}^i = \frac{1}{(c * \sigma_{RSTD})^2} * \begin{pmatrix} (\cos\theta_1 - \cos\theta_i)^2 & (\cos\theta_1 - \cos\theta_i)(\sin\theta_1 - \sin\theta_i) \\ (\cos\theta_1 - \cos\theta_i)(\sin\theta_1 - \sin\theta_i) & (\sin\theta_1 - \sin\theta_i)^2 \end{pmatrix} \quad (2.8)$$

θ_i is the angle of the line between UE and ith cell positive counter-clockwise from the X-axis. σ_{RSTD} is the RMS of RSTD measurement error. The detailed formula for RSTD is shown in [3].

TA can be used to estimate the distance from the UE to serving eNodeB. Information matrix of TA can be calculated by using the same principles as RSTD [3]:

$$I_{TA} = \frac{1}{(c * \sigma_{TA})^2} \begin{pmatrix} \cos^2\theta_{HS} & \cos\theta_{HS}\sin\theta_{HS} \\ \cos\theta_{HS}\sin\theta_{HS} & \sin^2\theta_{HS} \end{pmatrix} \quad (2.9)$$

σ_{TA} is the RMS of TA measurement error. The detailed formula for TA is shown in [2].

The variance of RSRP, RSTD and TA can be assumed to as σ_{RSRP}^2, σ_{RSTD}^2, σ_{TA}^2, respectively.

III. PROPOSED POSITIONING METHOD

Velocity can be obtained by assuming UE move from one point to another point during the certain time and it is expressed as a function of UE's coordinates information. It is useful to distinguish UE status if the velocity can be used appropriately, so the UE can be located well and truly. Based on current Cramer-Rao lower bound in section Ⅱ, the positioning performance can be re-evaluated considering the UE velocity in formula deducing. It can be expected that positioning error lower bound of new RFPM with velocity is smaller than traditional RFPM method.

Velocity can be calculated by the UE moving distance divided the corresponding time. It is assumed that UE move

 ISBN: 978-1-61208-203-5

from $dp_1 = \left(x_1^{dp1} \quad x_2^{dp1}\right)$ to $dp_2 = \left(x_1^{dp2} \quad x_2^{dp2}\right)$ during the time of Δt. So Velocity can be expressed as below:

$$v = \frac{\sqrt{\left(x_1^{dp1} - x_1^{dp2}\right)^2 + \left(x_2^{dp1} - x_2^{dp2}\right)^2}}{\Delta t} \tag{3.1}$$

So, differential coefficient of velocity can be obtained:

$$\frac{\partial v}{\partial dp_2} = \frac{1}{\Delta t}\left(\cos(\theta_v) \quad \sin(\theta_v)\right)$$

where

$$\cos(\theta_v) = \frac{\left(x_1^{dp2} - x_1^{dp1}\right)}{\sqrt{\left(x_1^{dp1} - x_1^{dp2}\right)^2 + \left(x_2^{dp1} - x_2^{dp2}\right)^2}}$$

$$\sin(\theta_v) = \frac{\left(x_2^{dp2} - x_2^{dp1}\right)}{\sqrt{\left(x_1^{dp1} - x_1^{dp2}\right)^2 + \left(x_2^{dp1} - x_2^{dp2}\right)^2}}$$

The information can be calculated as below:

$$I_v = \frac{1}{\sigma_v^2 * \Delta t^2}\begin{pmatrix} \cos^2(\theta_v) & \cos(\theta_v) * \sin(\theta_v) \\ \cos(\theta_v) * \sin(\theta_v) & \sin^2(\theta_v) \end{pmatrix} \tag{3.2}$$

θ_v is the angle of the line between UE last position and current position positive counter-clockwise from the X-axis.

σ_v is the RMS of velocity measurement error. We can calculate information matrix of this new positioning method by using (2.6), (2.8),(2.9), (3.2):

$$I = I_v + I_{RSRP} + I_{RSTD} + I_{TA} \tag{3.3}$$

So, error can be calculated by using (2.2) and (3.3):

$$\sigma_{LOC} = \sqrt{\text{trace}(I^{-1})} = \sqrt{\frac{M}{N}} \tag{3.4}$$

where

$$M = \frac{p^2 + q^2}{\sigma_{RSRP}^2} + \frac{1}{(c * \sigma_{TA})^2} + \frac{r}{(c * \sigma_{RSTD})^2} + \frac{1}{\sigma_v^2 * \Delta t^2}$$

where $r = \sum_{i=[2,N_{RSTD}]} (\cos\theta_1 - \cos\theta_i)^2 + \sum_{i=[2,N_{RSTD}]} (\sin\theta_1 - \sin\theta_i)^2$

N_{RSTD} is the number of cells participating the OTDOA positioning.

$$N = \frac{N_1}{\sigma_{RSRP}^2 * \sigma_{RSTD}^2 * c^2} + \frac{N_2}{\sigma_{TA}^2 * c^2 * \sigma_v^2 * \Delta t^2} + \frac{N_3}{\sigma_{TA}^2 * \sigma_{RSTD}^2 * c^4} + \frac{N_4}{\sigma_v^2 * \Delta t^2 * \sigma_{RSTD}^2 * c^2} + \frac{N_5}{\sigma_{RSRP}^2 * \sigma_{TA}^2 * c^2} + \frac{N_6}{\sigma_{RSRP}^2 * \sigma_v^2 * \Delta t^2} + \frac{N_7}{(c * \sigma_{RSTD})^4}$$

The related parameters used in above formulas can be summarized as follows:

$$N_1 = (pa_1 - qa_2)^2 \sum_{i=[2,N_{RSTD}]} (\rho_i)^2 + (pa_2 + qa_1)^2 \sum_{i=[2,N_{RSTD}]} (\lambda_i)^2 - 2 * (pa_2 + qa_1)(pa_1 - qa_2) \sum_{i=[2,N_{RSTD}]} (\rho_i \lambda_i)$$

$$N_2 = \left(\sin(\theta_{HS} - \theta_v)\right)^2$$

$$N_3 = \cos^2\theta_{HS} * \sum_{i=[2,N_{RSTD}]} (\rho_i)^2 + \sin^2\theta_{HS} * \sum_{i=[2,N_{RSTD}]} (\lambda_i)^2 - 2 * \cos\theta_{HS} \sin\theta_{HS} \sum_{i=[2,N_{RSTD}]} (\rho_i \lambda_i)$$

$$N_4 = \cos^2\theta_v * \sum_{i=[2,N_{RSTD}]} (\rho_i)^2 + \sin^2\theta_v * \sum_{i=[2,N_{RSTD}]} (\lambda_i)^2 - 2 * \cos\theta_v \sin\theta_v \sum_{i=[2,N_{RSTD}]} (\rho_i \lambda_i)$$

$$N_5 = ((pa_1 - qa_2) * \sin\theta_{HS} - (pa_2 + qa_1) * \cos\theta_{HS})^2$$

$$N_6 = ((pa_1 - qa_2) * \sin\theta_v - (pa_2 + qa_1) * \cos\theta_v)^2$$

$$N_7 = \sum_{i=[2,N_{RSTD}]} (\lambda_i)^2 * \sum_{i=[2,N_{RSTD}]} (\rho_i)^2 - \left(\sum_{i=[2,N_{RSTD}]} \lambda_i \rho_i\right)^2$$

where $\lambda_i = \cos\theta_1 - \cos\theta_i$ and $\rho_i = \sin\theta_1 - \sin\theta_i$.

IV. Simulation Results Analysis

To verify the proposed scheme, we simulated both traditional RFPM and our strategy in Urban environment scenario. In this section, the simulator of RFPM with Velocity is described in detail.

The 19 (sites) * 3 (sectors) topology is used in the Fig. 1, and the evaluation area is covered by the 3 center green sectors where UEs are uniformly dropped in our simulator. The site indexes are illustrated, and the bound of sectors is denoted by the dash line [3].

Inter-site distance (ISD) is 500 meter. Pathloss exponent is 0.5 and FBR equals to 30dB. Reference distance can be set 100 meter so that reference power can be calculated as -80.5 dBm according to hata model. Minimum Distance between target UE and eNodeB is 35m. The RSRP measurement error is defined in the RSRP, RSTD and TA accuracy requirement of [4].

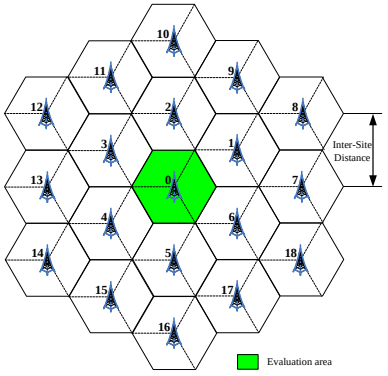


Figure 1. Network topology

For our simulation, the ±8 dB measurement error is assumed, while the RSTD and TA measurement error is assumed to be 5Ts, 10Ts respectively. Ts is the minimum time resolution of LTE systems, which equals to 1/(2048*15000)(s). Velocity variance is equipment implementation specific, and according to existing data, both1m/s and 0.5m/s are reasonable error value for this simulation. Δt can be calculated by using the equation as below:

$$\Delta t = \frac{ISD}{v} \tag{4.1}$$

In our simulation, we make a comparison between our strategy and traditional RFPM, in which Velocity is assumed to be 3km/h and 120km/h respectively. All the detailed simulation parameters are listed in Table I.

TABLE I. SIMULATION ASSUMPTIONS

Parameters	Value
Inter-site distance	512 m
Reference distance	100 m
FBR	30 dB
Reference power	-80.5 dBm
RSRP measurement errors	10 dB
TA measurement errors	10Ts
RSTD measurement errors	5Ts
Velocity variance	1m/s, 0.5m/s
Minimum distance between target UE and eNodeB	35 m
Pathloss exponent	0.5
N_{RSTD}	4
UE velocity	3/120 km/h

RFPM is related with the number of measurement elements. It means that if velocity is used for mapping the existing RFPM performance can be enhanced from theory aspect.

Based on the assumptions and Cramer-Rao lower Bound deducing, the simulation results can be obtained as shown in Figure 2 and Figure 3, improvement is quite obvious. According to the simulation results figures, the following table can be achieved. Comparison of two positioning methods performance is given in Table II.

TABLE II. COMPARISON RESULTS SUMMARY

Velocity Variance	1m/s	0.5m/s
Positioning error		
RFPM with 3km/h	49.6m	47.2m
RFPM with 120km/h	46.6m	45.7m
RFPM without velocity	66.9m	66.9m

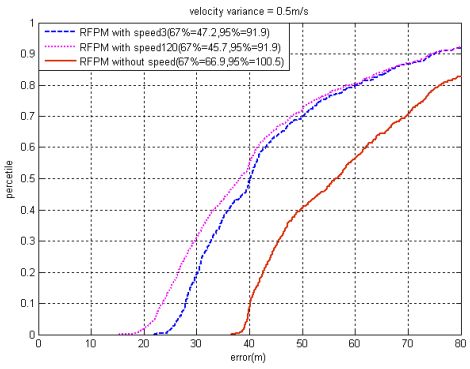


Figure 2. Comparison of methods (0.5m/s)

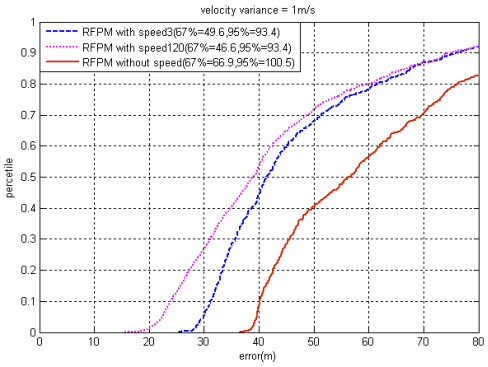


Figure 3. Comparison of methods (1m/s)

V. CONCLUSION AND FUTURE WORKS

In this paper, several positioning error models are provided based on the current 3GPP RAN4 protocols and the Cramer-Rao lower bound is used for positioning accuracy performance evaluation. As a novel update to the conventional schemes, the UE velocity is adopted to improve the RF pattern matching efficiency, and this novel RFPM scheme can remarkably enhance the UE positioning accuracy with 25.8%~32.2% gain compared with the traditional RFPM from Cramer-Rao lower bound deducing and related simulations. Besides the velocity information, other signature information can also improve the RFPM performance, such as temperature information, and these signature information for RFPM can be included in future works.

ACKNOWLEDGMENT

This work derived from cooperation project of Huawei Technologies Co., Ltd and Beijing University of Posts and Telecommunications.

REFERENCES

[1] FCC, “Revision of the Commission’s Rules to Ensure Compatibility with Enhanced 911 Emergency Calling Systems,” CC Docket 94-102, Jul. 1996.

[2] 3GPP TSG-RAN WG4, “Content for TR 36.809 (Study on the inclusion of RF Pattern Matching Technologies as a location method in the E-UTRAN)”, R4e-110006, Polaris wireless, Apr. 2011.

[3] 3GPP TSG-RAN WG4, “Discussion on RF pattern matching for positioning”, R4-114510, Huawei, HiSilicon, Aug. 2011.

[4] 3GPP, TS36.133 v10.6.0, “Requirements for support of radio resource management”, Mar. 2012.

[5] Xingchuan Liu, Sheng Zhang, Qingyuan Zhao, and Xiaokang Lin, "A novel approach for fingerprint positioning based on spatial diversity," Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on , vol. 6, no., pp. V6-441-V6-445, 20-22 Aug. 2010.

[6] Takenga, C., Tao Peng, and Kyamakya, K., "Post-processing of Fingerprint Localization using Kalman Filter and Map-matching Techniques," Advanced Communication Technology, The 9th International Conference on , vol. 3, no., pp. 2029-2034, 12-14 Feb. 2007

[7] Takenga, C. and Kyamakya, K., "A Low-cost Fingerprint Positioning System in Cellular Networks," Communications and Networking in China, 2007. CHINACOM '07. Second International Conference on, vol., no., pp.915-920, 22-24 Aug. 2007

 ISBN: 978-1-61208-203-5

Enabling Guaranteed Beacon and Data Slots in Multi-hop Mesh Sensor Networks for Home Health Monitoring

Juan Lu[1,2], Adrien Van Den Bossche[1,3] , Eric Campo[1,2]
[1]Université de Toulouse; UPS, INSA, INP, ISAE, UTM, F-31703 Blagnac, France
[2]CNRS; LAAS; 7 Avenue du Colonel Roche, F-31077 Toulouse, France
[3]CNRS; IRIT; 118 Route de Narbonne, F-31062 Toulouse, France
{lu, vandenbo, campo}@iut-blagnac.fr

***Abstract*—The MAC (Medium Access Control) protocol has an important influence on network performance, especially in home health monitoring which constrains delivery of time-sensitive message and power consumption. In this paper, a multi-hop mesh sensor network is proposed based on a novel MAC protocol ADCF (Adaptive and Distributed Collision-Free). ADCF uses CFBS (Collision-Free Beacon Slot) and CFDS (Collision-Free Data Slot) mechanisms to guarantee QoS (Quality of Service) while reducing energy consumption in a mesh topology. The simulation results show better performances of ADCF compared with IEEE 802.15.4 MAC protocol in terms of energy and guaranteed medium access with the flexibility of mesh topology.**

Keywords-IEEE 802.15.4; mesh topology; QoS; energy saving, health monitoring application

I. INTRODUCTION

Our application scenario is focused on the monitoring of the elderly at home via a WSN (Wireless Sensor Network). ADCF sensor nodes are put on the ceiling, wall, furniture and the body of the person. For example, when the accelerometer detects a fall of the person, an alarm should be sent with some guaranties in terms of delay. In addition, the network should be self-organizing and could tolerate a link failure or a link establishment. Therefore, all ADCF nodes are expected to have the same role (both sensor and router) in a mesh topology.

Several projects have been investigated on the habitat monitoring [1-3]. There are generally two main constraints for this WSN: time-sensitive delivery of some urgent messages and power consumption. Many technologies exist at different layers to improve these two constraints [4]. Our work focuses on MAC layer. As the largest energy consumption of the nodes is due to the time spent in the idle state [5], so time slot allocation is an important task. The avoidance of collisions between 2-hop neighbors is another goal because there is scarcely interference at distance of more than 2 hops [6].

This paper aims to present a novel MAC protocol based on IEEE 802.15.4 to build a scalable and robust WSN for home health monitoring. The paper is organized as follows: section 2 investigates the current MAC protocol for this application. The proposed ADCF MAC is described gradually in section 3. Section 4 provides simulation results while the last section concludes the paper.

II. RELATED WORK

MAC protocols for WSN could be classified into two categories. The first category is based on conventional wireless protocols, especially IEEE 802.11(a/b/g/n/ac). These protocols typically provide a general mechanism that works reasonably well for a large set of traffic load. Therefore, these protocols don't meet our goals and will not be discussed in this paper. IEEE 802.11ah is an on-going work about energy efficient MAC for low traffic sensor network. However, some issues such as frame header compression are still open and there are now no available products for our future work. The second category based on IEEE 802.15.4 is being considered as a promising way for low-cost low-power WSN. In part A, IEEE 802.15.4 standard is briefly presented. Recent works based on this standard have been fully studied in part B.

A. IEEE 802.15.4 Standard

IEEE 802.15.4 [7] protocol supports beacon and non-beacon mode. More precisely, in beacon mode, it is possible to achieve variable duty cycles (from 100% down to 0.006%), which is particularly interesting for our application where energy constraint and network lifetime are main concerns. In addition, beacon mode has an attractive feature for time-sensitive applications as QoS properties are available with GTS (Guaranteed Time Slot) mechanism. On the other side, non-beacon mode, which has the advantage of lower complexity and more scalability as compared with beacon mode, does not provide any of those features.

Therefore, we focus on beacon mode which seems to be a promising way. However, several issues in the standard are still open. One of those issues is how to build a synchronized multi-hop mesh network for power efficient, scalable and robust networking. In fact, while the current standard supports multi-hop networking using peer-to-peer topology, it restricts its use to non-beacon mode. This contradiction makes urgent requirement of novel MAC protocols.

 ISBN: 978-1-61208-203-5

B. MAC Protocols Based on IEEE 802.15.4

ZigBee specifications [8] clear the ambiguities of IEEE 802.15.4 in a cluster-tree topology. The centralized PAN (Personal Area Network) coordinator assigns a beacon transmission offset for each node when it wants to associate the PAN. Therefore, the communication range and the requirements of time-sensitive are both limited.

Anis Koubâa [9-10] continues the work in the domain of cluster-tree topology. However, the requirement of different BI (Beacon Interval) and SD (Superframe Duration) for each node is calculated in advance. These weaken the flexibility and robustness as well as restrict the scalability of network.

Another example has been proposed in OCARI project [11-12]. A PAN coordinator which receives all the association requests decides beacon slot for each node. The main drawback of this solution is the lack of flexibility, especially regarding the changing topology and the inconstancy of wireless medium.

P. S. Muthukumaran presented MeshMAC protocol [13]. This protocol enables mesh networking over beacon mode through a distributed SDS (Superframe Duration Scheduling) strategy in which each node calculates its schedule to transmit beacons based only on locally available information. The limitations of MeshMAC are: it imposes very low duty cycles; the beacon transmission offset is difficult to choose for the changing topology.

B. Carballido Villaverde proposed DBOP MAC protocol [14]. It creates a BOP (Beacon Only Period) where beacons are transmitted at different time slots among neighbors and neighbors' neighbors. However, DBOP introduces an overhead into the network. Another drawback is the inefficient management of BOP length. In addition, how to offer QoS for different application traffic is not discussed.

III. ADCF MAC PROTOCOL

The objective of ADCF is to build a beacon-enabled WSN over an IEEE 802.15.4 PHY which supports mesh topology and enables better energy efficiency. While a previous paper [15] only focused on beacon scheduling and network construction, in this paper we detail the mechanism of Collision-Free Data Slot (CFDS) in the mesh topology. Additionally, corresponding simulation results and the operation of ADCF are first presented.

Before showing the characteristics of ADCF, some assumptions should be highlighted: all the considered nodes have the capacity to be both sensor and router; nodes addresses have been preliminary set.

A. Overview of ADCF

As shown in Fig. 1, the superframe of ADCF is organized in three parts: BOP, active period and inactive period. BOP is organized by CFBS (Collision Free Beacon Slot). Each node has a 2-hop collision-free beacon slot in BOP. Similarly, ADCF nodes can access the medium by slotted CSMA or guaranteed mechanism CFDS (Collision Free Data Slot) in the active period. Inactive period is optional for energy saving.

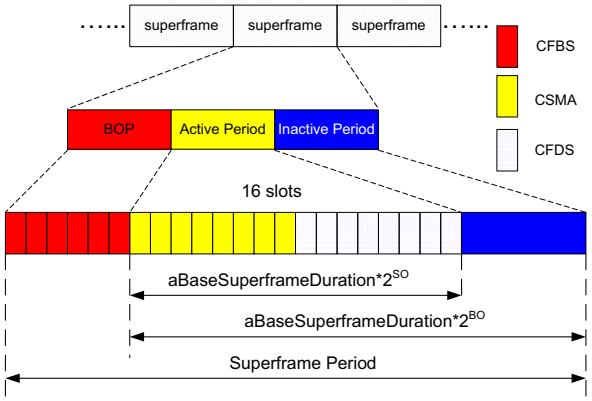


Figure 1. ADCF superframe structure.

The basic parameters are consistent with IEEE 802.15.4:

- Active period is divided into 16 slots.
- 0 ≤ SO (Superframe Order) ≤ BO (Beacon Order) ≤14.
- aBaseSuperframeDuration denotes the number of symbols that form a superframe when SO is 0.

B. Operation of ADCF

ADCF includes several slight protocols: BEP (Beacon Exchange Protocol), ISP (Initiator Selection Protocol), BSAP (Beacon Slot Allocation Protocol), DSAP (Data Slot Allocation Protocol) and SRP (Smart Repair Protocol). In addition, SPA (Simple Priority Algorithm) is used repeatedly in ISP and BSAP.

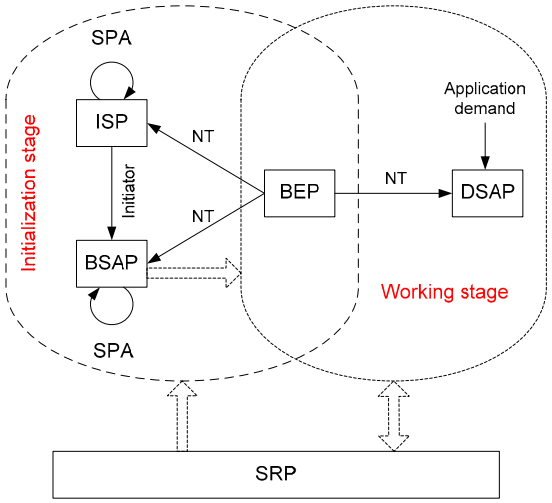


Figure 2. ADCF operation diagram.

Each ADCF node has a NT (Neighbor Table) and executes the following as shown in Fig. 2. SRP allows

ADCF nodes to switch between initialization stage and working stage depending on the changing topology of network. The beginning and core of ADCF is BEP which sets up NT and updates NT in both stages. BEP runs periodically according to the preset parameters such as BO. With the information of NT, ISP is executed. Then BSAP is triggered and so the node could synchronize with the initiator. DSAP will work when there are application requests from the higher layer.

- **BEP**: the main concern for BEP is collection of interesting information in a 2-hop neighborhood. Each new node will firstly listen to the channel for a fixed period when it is powered-up. Depending on the received beacons during listen, the new node will send its own beacon by different mechanisms. Each node broadcasts its beacon within 1-hop and records direct neighbors' in its NT. Therefore, all the 2-hop neighbors' information is obtained by this new node. The interesting information in a beacon includes NA (Neighbor Address), NE (Neighbor Energy) and ND (Neighbor Density). Here, ND is defined as the number of neighbors within 2-hop (including itself). The overhead incurred by BEP is studied and simulated in [15].
- **SPA**: SPA is implemented by comparing 3 parameters of the nodes. The comparison order is ND, NE and NA. At first, the node with maximum ND is selected. If the nodes have the same ND, SPA chooses the one with maximum NE. Finally, the node with minimum address has the highest priority if two other parameters are the same.
- **ISP**: the objective of this protocol is to select an initiator which has two functions: it specifies the beginning of BOP and measures the length of BOP in order to realize the network synchronization. This length is defined as the initiator ND. Each node selects an initiator candidate locally by SPA from its NT. If one initiator candidate is different from the neighbors', SPA is repeatedly used to decide a unique initiator. This initiator's information will be added to NT and be sent in the next beacon. Therefore, there may be several initiator candidates in the initialization stage but a unique initiator in the working stage.
- **BSAP**: this protocol makes each node choose a CFBS in BOP. The nodes execute SPA locally and the one has higher priority first to choose its beacon slot. It takes a slot which is not used by its 2-hop neighbors and stores the slot number in its NT. At last, the node which has its chosen slot will be deleted from SPA list and the other nodes in this list continue BSAP.
- **DSAP**: in the original IEEE 802.15.4, GTSs are requested via a GTS request command sent in Best-effort mode using CSMA/CA. In ADCF, each node can request CFDS using its beacon to all its neighbors without the need to send a dedicated frame. A bi-direction communication is possible. When a node receives neighbor's beacon and finds its address as CFDS destination, it checks it NT, allocates the first available data slot to the requesting node and announces this allocation in the next beacon. When the requesting node receives this beacon containing the slot number, it may use it to send the application traffic in the CFDS. CFDS request and CFDS indication subfields take only 2 bytes in the beacons.
- **SRP**: this protocol reduces the impact of changing topology as much as possible. For example, if link failure is detected, neighbors will simply delete this failure node from their NT. If the initiator fails, others re-select an initiator but keep their BOP with the original beacon slots. Therefore, the network will still work without disruption. As there are free slots in BOP, a new node may choose its beacon slot directly after a period of listening. If BOP length is not enough, a new node may send its beacon by CSMA until a new initiator is designed with updated ND.

In conclusion, seldom MAC protocols for low-cost low-power network are in a mesh topology which has the advantages of scalability and robustness. ADCF aims to enable the efficient mechanisms and eliminate the difficulties, such as beacon collision, QoS and synchronization in a changing multi-hop mesh-link network.

IV. SIMULATION STUDY

To study the scope of our contribution, we use OPNET to establish a simulation model which implements the entire proposal. Two experiment examples are presented in this paper. The first one is the comparison of ADCF with IEEE 802.15.4. The second experiment is the ADCF performance with large scale and high neighbor density. The basic parameters are shown in the Table 1. We are now implementing ADCF on 13192-SARD board which has a total of 4Kb RAM for application data, variables, buffers etc. Therefore, the buffer for CSMA and CFDS could not be more than 2Kb in reality. This parameter configuration is useful for comparing the simulation results with prototype.

TABLE I. TABLE TYPE STYLES

Parameter	Value
Scene area	100*100 m^2
Transmission range	15 m
BO	7
SO	4
Traffic distribution	Constant
Application payload	100 bits
CSMA buffer	0.5 k octets
CFDS buffer	1.5 k octets
Simulation duration	30 min
Simulation times	20

A. Comparison of ADCF with IEEE 802.15.4

To our knowledge, there are two versions of IEEE 802.15.4 in OPNET. The version developed by Anis Koubâa [9-10] includes the GTS implementation and a fixed beacon

scheduling mechanism. It is used in this experiment. 14 nodes join the network gradually in this experiment. A static routing mechanism is added above ADCF in order to simulate real traffics over the network. IEEE 802.15.4 applies ZigBee routing.

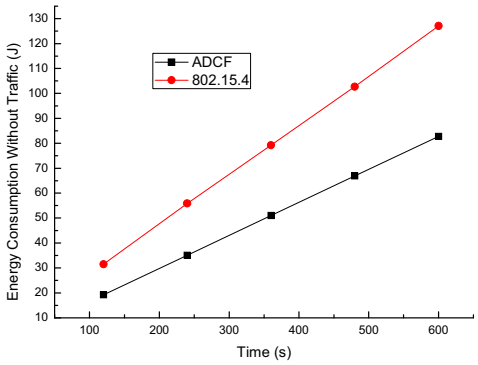


Figure 3. Energy consumption comparison.

In the energy consumption comparison, only beacons are delivered in order to compare the protocol cost. In addition, a more practical energy model [16] is used in our experiment. As shown in Fig. 3, ADCF consumes less energy, about 37%, than IEEE 802.15.4 as time goes by. This is because IEEE 802.15.4 nodes spend more time for idle listening.

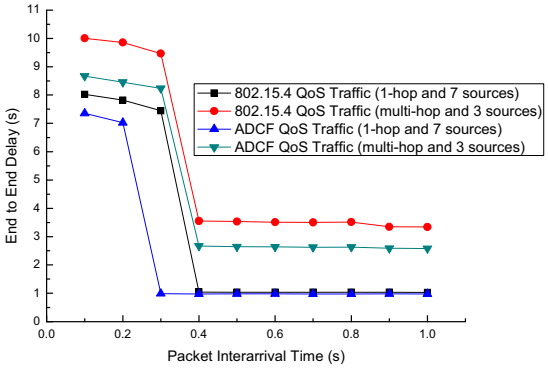


Figure 4. Delay comparison.

As shown in Fig. 4, there are 7 sources with 1-hop traffic or 3 sources with multi-hop traffic. When Packet Interarrival Time decreases from 1.0 to 0.1, the traffic load will increase. Therefore, End to End Delay becomes larger. When Packet Interarrival Time is about 0.4 s, there are radical changes caused by buffer overflow.

As the configured BO and SO, a superframe cycle is about 2 s. For 1-hop traffic, the difference between ADCF and IEEE 802.15.4 is tiny. This End to End Delay, about 1 s, includes the time from packet generation to the scheduled data slot. It also can be seen that ADCF saves about 25% End to End Delay for multi-hop traffic. Some multi-hop traffic may be transmitted in one superframe as the same active period in a mesh topology. IEEE 802.15.4 works in a cluster-tree topology which may take several superframes from the source to the final destination. The average hop count is 3, so this End to End Delay is about 2.7 s for ADCF and 3.5 s for IEEE 802.15.4.

The Packet Success Ratios always keep 100% for both protocols when the CFDS buffers are available.

B. ADCF Performance in Large Scale and High Density

In this experiment, we focus on Packet Success Ratio of ADCF in large scale and high density in order to study the protocol performance in a variety of scenarios.

Firstly 14, 30 and 50 nodes are configured in the network. Then the 50 nodes with different neighbor density are simulated. For all the scenarios, there are 7 sources with QoS traffic and 7 sources with best-effort traffic at the same time. All traffics are generated for a 1-hop destination.

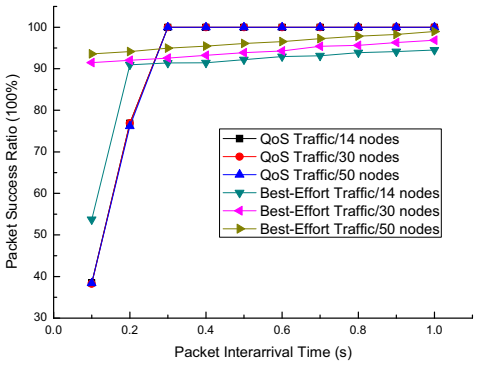


Figure 5. Packet success ratio for different scale.

As shown in Fig. 5, it can be seen that Packet Success Ratios always keep 100% for QoS traffic when the CFDS buffers are available. Packet Success Ratios become higher with the larger network scale for Best-Effort traffic. This is because of the risks of collisions are lower. When Packet Interarrival Time is 0.1, this traffic load is relatively light for the network of 30 and 50 nodes. However, the contention for Best-Effort traffic is intense in the network of 14 nodes.

Neighbor densities are average values obtained by simulations. As shown in Fig. 6, network density also has no much influence on QoS traffic. While for Best-Effort traffic, Packet Success Ratio is higher with the lower density as there is less collisions. When neighbor density is 8.76, the risk of collision is low. Therefore, the difference between 8.76 and 5.13 is tiny. When neighbor density is 15.24, there are a lot of nodes in the communication range of neighbors. Thus its Packet Success Ratio is the lowest of these 3 scenarios.

 ISBN: 978-1-61208-203-5

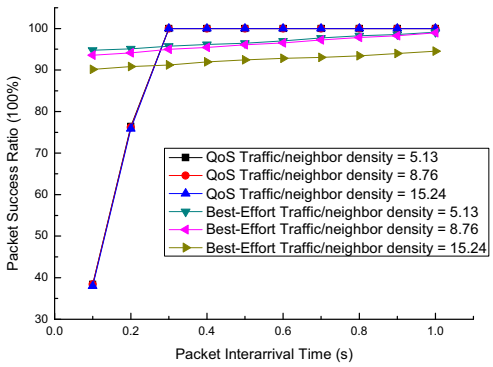


Figure 6. Packet success ratio for different density.

V. CONCLUSION AND PERSPECTIVES

This paper presents an original MAC protocol named as ADCF, which is based on IEEE 802.15.4 standard to build a mesh WSN. The 2-hop CFBS and CFDS mechanisms were described and implemented by a set of protocols which are explained. The simulation results show that ADCF consumes less energy (about 37%) while End to End Delay and Packet Success Ratio perform no much worse than IEEE 802.15.4. The simulation results also confirm that ADCF works well in the condition of large scale and high density. Therefore, ADCF satisfies the application request of delivering QoS message with low energy consumption. In addition, when a new node joins the network or a key node fails in the process of surveillance, the own functioning of other nodes is quite important for home health monitoring. Fortunately, the mesh topology of ADCF strengthens network flexibility to the changing link states. A perspective is the re-exploitation of the information gathered by ADCF in order to make them available for upper layers, such as routing layer, to reduce upper protocol overhead.

Now, the current work is focused on ADCF hardware implementation. The next step is its deployment in real conditions in "Smart Home" of Blagnac University Technological Institute.

ACKNOWLEDGMENT

This work is partly supported by the Academic OPNET Research and Educational Projects. Authors promptly acknowledge the support.

REFERENCES

[1] S. Nourizadeh, C. Deroussent, Y. Q. Song, J. P. Thomesse, "Medical and Home Automation Sensor Networks for Senior Citizens Telehomecare", IEEE International Conference on Communications (ICC Workshops 09), Jun. 2009, pp. 1-5.

[2] Hongwei Huo, Youzhi Xu, Hairong Yan, Saad Mubeen, Hongke Zhang, "An Elderly Health Care System Using Wireless Sensor Networks at Home", Third International Conference on Sensor Technologies and Applications (SENSORCOMM 09), Jun. 2009, pp. 158-163.

[3] Y. Zatout, E. Campo, J. P. Llibre, "WSN-HM: Energy-Efficient Wireless Sensor Network for Home Monitoring", Fifth International Conference on Intelligent Sensors Sensor Network and Information Processing (ISSNIP 09), Dec. 2009, pp. 367-372.

[4] A. H. Salem, A. Kumar, A. S. Elmaghraby, "A Survey for QoS and Power Management Algorithms in Wireless Transmission", IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 06), Aug. 2006, pp. 720-727.

[5] S. Mahfoudh, P. Minet, "Maximization of Energy Efficiency in Wireless Ad hoc and Sensor Networks with SERENA", Mobile Information Systems, Advances in Wireless Networks, Volume 5 Issue 1, Apr. 2009, pp. 33-52.

[6] A. van den Bossche, T. Val, E. Campo, "Prototyping and performance analysis of a QoS MAC layer for industrial wireless network", 7th International Conference on Fieldbuses and nETworks in industrial and embedded systems (IFAC 07), Nov. 2007, Volume 7 Part 1.

[7] IEEE 802.15.4 Standard (2006) Part 15.4: Wireless medium access control (MAC) and physical layer (PHY) specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs), IEEE Standard for Information Technology, IEEE-SA Standards Board.

[8] ZigBee-Alliance,http://www.zinbee.org.

[9] Anis Koubaa, André Cunha, Mário Alves, "A Time Division Beacon Scheduling Mechanism for IEEE 802.15.4/Zigbee Cluster-Tree Wireless Sensor Networks", 19th Euromicro Conference on Real-Time Systems (ECRTS 07), Jul. 2007, pp. 125-135.

[10] A. Koubâa, M. Alves, M. Attia, A. Van Nieuwenhuyse, "Collision-Free Beacon Scheduling Mechanisms for IEEE 802.15.4/Zigbee Cluster-Tree Wireless Sensor Networks", 7th International Workshop on Applications and Services in Wireless Networks (ASWN 07), May 2007.

[11] T. Dang, C. Devic, E. Livolant, A. Van Den Bossche, T. Val, "OCARI: Optimization of Communication for Ad Hoc Reliable Industrial Networks", 6th IEEE International Conference on Industrial Informatics (INDIN 08), Jul. 2008, pp. 688-693.

[12] K. Alagha, G. Chalhoub, A. Guitton, E. Livolant, S. Mahfoudh, P. Minet, M. Misson, J. Rahme, T. Val, A. van den Bossche, "Cross-Layering in an Industrial Wireless Sensor Network: Case Study of OCARI", Journal of Networks, Vol.4 issue 6, Aug. 2009, pp. 411-420.

[13] P. S. Muthukumaran, R. de Paz, R. Špinar, and D. Pesch, "MeshMAC: Enabling Mesh Networking over IEEE802.15.4 through Distributed Beacon Scheduling", AD HOC Networks, Vol. 28 Part 1, Jan. 2010. pp. 561–575.

[14] B. Carballido Villaverde, R. De Paz Alberola, S. Rea, D. Pesch, "Experimental Evaluation of Beacon Scheduling Mechanisms for Multihop IEEE 802.15.4 Wireless Sensor Networks", 4th Conference on Sensor Technologies and Applications (SENSORCOMM 10), Jul. 2010, pp. 226-231.

[15] J. Lu, A. van den Bossche, E. Campo, "An Adaptive and Distributed Collision-Free MAC Protocol for Wireless Personal Area Networks", 6th International Symposium on Intelligent Systems Techniques for Ad hoc and Wireless Sensor Networks (IST-AWSN 11), Sep. 2011, pp. 798-803.

[16] N. Fourty, A. van den Bossche, T. Val, "Etude de l'impact énergétique de l'algorithme d'accès au médium pour un réseau de capteurs sans fil industriel", Sixième Conférence Internationale Francophone d'Automatique (CIFA 10), Jun. 2010.

 ISBN: 978-1-61208-203-5

AEGIR – Asynchronous Radiolocation System

Slawomir J. Ambroziak, Ryszard J. Katulski, Jaroslaw Sadowski, Wojciech Siwicki, Jacek Stefanski
Department of Radiocommunication Systems and Networks
Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics
Narutowicza Street 11/12, 80-233 Gdansk, Poland
{sj_ambroziak, rjkat, jaroslaw.sadowski, wojciech.siwicki, jstef}@eti.pg.gda.pl

Abstract— **Humans have always wanted to determine position in an unknown environment. At the beginning methods were simple. They were based on the observation of characteristic points, in the case of shipping additional observations of the coastline. Then came navigation based on astronomical methods (astronavigation). At the beginning of the XX-century a new way of determining the current location was developed. It has used radiowave signals. First came radio-beacons. Then ground-based systems came. Currently satellite systems are being used. At present, the most popular one is Global Positioning System (GPS). This system is fully controlled by the Department of Defense, and only the U.S. forces and their closest allies have been guaranteed accuracy offered by the system. Armies of other countries can only use the civilian version. This situation has engendered the need for an independent radiolocation system. This article describes the construction and operation of such a technology demonstrator that was developed at Gdansk University of Technology. It was named AEGIR (according to Norse mythology: god of the seas and oceans). The main advantage of the system is managing without the chain organization of the reference stations, which work now with each other asynchronously. This article demonstrates the functionality of such system. It also presents results and analysis of its effectiveness.**

Keywords- navigation; hyperbolic systems; radiolocation; AEGIR; TDOA.

I. INTRODUCTION

Global Navigation Satellite System (GNSS) is seen by terrorists or hostile countries as a high value target. Volpe Center report contains the following statement [1]: "During the course of its development for military use and more recent extension to many civilian uses, vulnerabilities of Global Navigation Satellite Systems (GNSS) – in the United States the Global Positioning System (GPS) – have become apparent. The vulnerabilities arise from natural, intentional, and unintentional sources. Increasing civilian and military reliance on GNSS brings with it a vital need to identify the critical vulnerabilities to civilian users, and to develop a plan to mitigate these vulnerabilities.". GNSS can also be targeted by more common criminals - computer hackers and virus writers. Therefore, there is a need for maintenance and continued development of independent radionavigation and radiolocation systems.

Based on many years of experience in the field of modern radiocommunication systems in the Department of Radiocommunication Systems and Networks at Gdansk University of Technology, in cooperation with the OBR Marine Technology Centre in Gdynia and with the support of the Hydrographic Office of Polish Navy a ground-based radiolocation system, which was named AEGIR has been developed, built and tested in real environment. In this system, all reference stations are working in an asynchronous way, so each station uses a local generator to transmit a message location and can receive signals from neighboring stations. On the basis of the received signals reference station determines the time difference between its own rhythm of work, and the neighboring reference stations. The measurement results are periodically placed in the localization message. The receiver on the basis of self-measurements and measurements from the reference stations estimates its location. Compared to existing solutions like Loran-C (Long Range Navigation - C) [2], the AEGIR system resigns chain relationship between reference stations. In the proposed system, there are no supervision centers for maintenance which reduces operating costs and increases system reliability. With this approach, our system has gained new features and new functionality compared to traditional solutions.

This paper at the beginning will present basics of the TDOA method. Then principle of asynchronous system will be described. The next section describes hardware implementation of the presented system. The last two describes investigation results and a brief summary.

II. HYPERBOLIC SYSTEMS – TDOA METHOD

As mentioned before, the AEGIR system is a ground based radiolocation system. Therefore a measurement method of Time Difference Of Arrival (TDOA) was chosen to estimate the position of a localizer. Suppose there are N ground stations, the coordinates for the i-th station are $S_i = (x_{Si}, y_{Si})$, where i = 1, ..., N, and the search object's coordinates are $M = (x_M, y_M)$.

If you define a signal propagation time between the i-th station and the searched position in the point M as T_i, so the distance between the i-th station and the point M is as follow:

 ISBN: 978-1-61208-203-5

$$d_i = T_i \cdot c = \sqrt{(x_{Si} - x_M)^2 + (y_{Si} - y_M)^2}, \quad (1)$$

where:

c - velocity of wave propagation (3 * 10^8 m / s)
T_i - the propagation delay between the i-th station and the point M,
d_i - distance between i-th station and the point M.

Timing differences between the i-th station and the first one, can be written as:

$$T_{i1} = T_i - T_1, , \quad (2)$$

Differences in the distances between those stations, can be described by the following relationship:

$$d_{i1} = T_{i1} \cdot c = d_i - d_1, \quad (3)$$

After putting equation (1) in equation (3) we obtain hyperbolic equation:

$$d_{i1} = \sqrt{(x_{Si} - x_M)^2 + (y_{Si} - y_M)^2} - \sqrt{(x_{S1} - x_M)^2 + (y_{S1} - y_M)^2}. \quad (4)$$

Equation 4 presents the difference in distance between the first and i-th station.

Determination of the distance difference between another pair of base stations generates more hyperbolas and a point of their intersection gives us a position. There are many algorithms [3-6], which allow to determine the coordinates. For the purpose of the system the Chan method was chosen [3], because it gives results without iterative calculations and additionally it is simple to implement.

III. ASYNCHRONOUS SYSTEM

As mentioned in the introduction, the AEGIR system is fully asynchronous. The principle of asynchronous method can be illustrated as follows. Assume that we have three reference stations positioned as in Fig. 1.

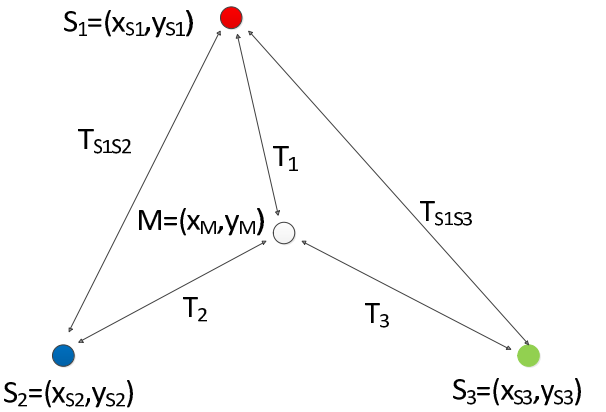


Fig. 1. Deployment of ground stations to illustrate the method of TDOA

Propagation time from the stations S_1, S_2 and S_3 to desired position in the point M (localizer) is respectively T_1, T_2 and T_3. Each station has coordinates as follows: S_1=(x_{S1}, y_{S1}), S_2=(x_{S2}, y_{S2}) and S_3=(x_{S3}, y_{S3}). Stations transmit a reference signal, for simplicity, as an impulse, but time of broadcasting these impulses, as shown in Fig. 2a, is random. The stations have the ability to "listen to" neighboring stations. This is illustrated in Fig. 2b. Reference station designated as S1 receives signal from other two stations: S2 and S3, and calculates the time difference between its own and these stations' signals (nT_{21} and nT_{31}). These time differences are then sent to the localizer. The localizer (at point M) (pictured in Fig. 2c) sets its own time difference between the received impulses from the reference station (dT_{21} and dT_{31}).

Additionally, each ground station sends to the localizer its own coordinates (respectively x_{S1}, y_{S1} - the coordinates of the first station; x_{S2}, y_{S2} - coordinates of the second station; x_{S3} and y_{S3} - coordinates of the third station), so that the localizer calculates the propagation time between the reference stations (T_{S1S2}, T_{S1S3}).

Taking into account all sent data, the localizer (at point M) calculates a real difference in time propagation between stations, which present the following equation:

$$T_{21} = nT_{21} - dT_{21} - T_{S1S2}, \quad T_{31} = nT_{31} - dT_{31} - T_{S1S3}, \quad (5)$$

The time differences defined in this manner allow to determine coordinates of searched object M using one of the algorithms [3-6].

 ISBN: 978-1-61208-203-5

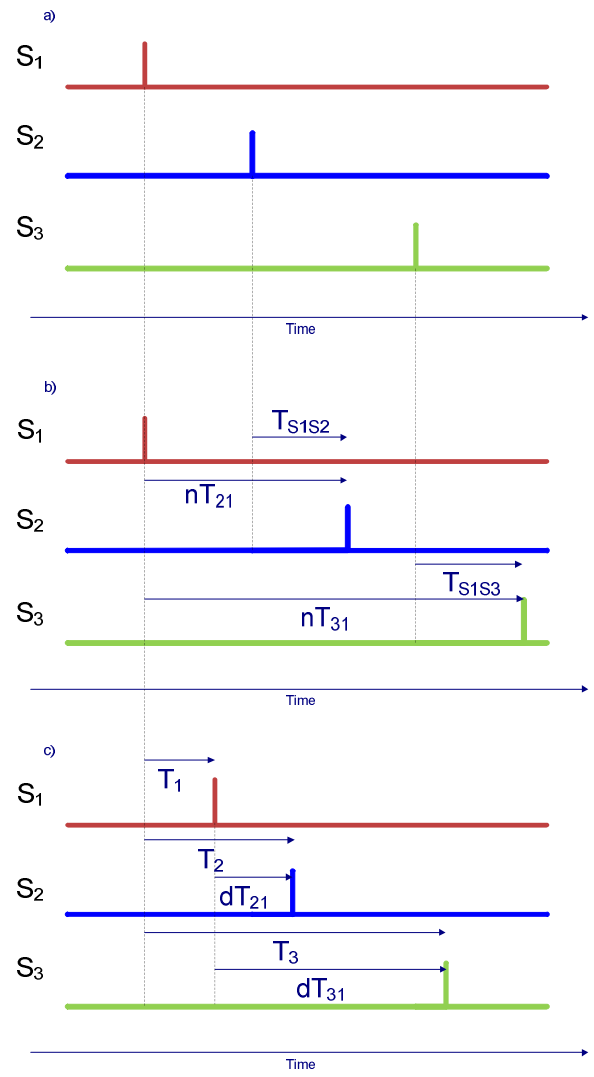


Fig. 2. Timing between signals broadcasted by base stations in an asynchronous system, a) the moment of broadcasting impulses by the stations b) the time of receipt of impulses by S_1 station c) the time of receipt of impulses by the localizer in point M [8,9]

IV. HARDWARE IMPLEMENTATION

The AEGIR system has been built as a demonstrator of technology. The system consists of a localizer and three reference stations.

The block diagram of a localizer is presented in Fig. 3.

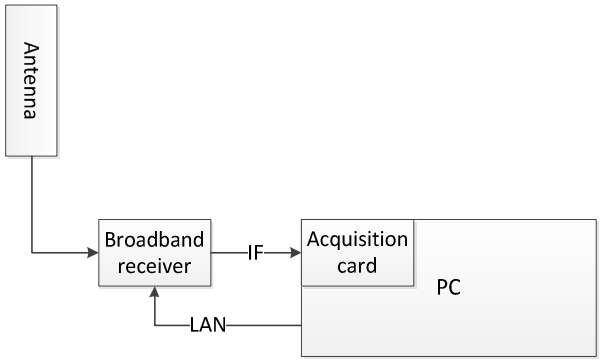


Fig. 3. Block diagram of a localizer.

The localizer has been made in the technology of Software Defined Radio [7]. It consists of: an antenna, a broadband receiver, an analog to digital converter (in the form of data acquisition card) and a digital signal processor (in form of PC). This approach allows to shape flexibly functionality of the localizer.

Ground stations, as it was mentioned before, have the ability to "listen to" neighboring stations. It is assumed that the system should consists only of such stations (Master ones). However, for demonstrable purposes only one Master station is required. Therefore, two types of ground stations were created: broadcasting stations (Slave type) and broadcasting and listening ones (Master type).

The block diagram of a Slave station is shown in Fig 4.

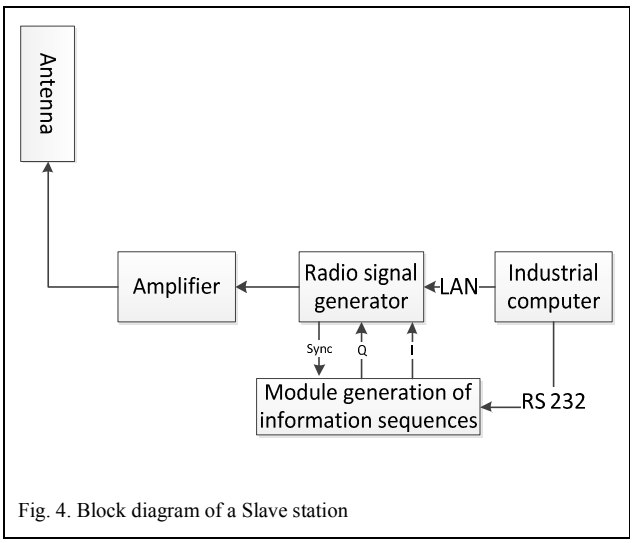


Fig. 4. Block diagram of a Slave station

The main element of the station is a radio signal generator, whose task is to broadcast modulated signal with data that are generated by industrial computer.

 ISBN: 978-1-61208-203-5

The block diagram of the last element of the described system - Master station – is shown in Fig. 5.

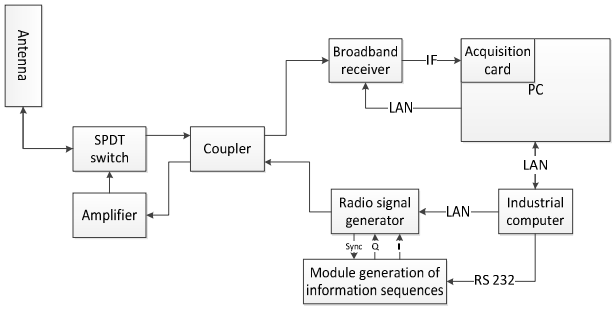


Fig. 5. Block diagram of a Master station

Master station is a combination of a localizer and a Slave station. The task of the receiver is to listen to a nearby station and to determine difference in synchronization between reference signal and signals from the neighboring stations. To enable listening to neighboring stations, Master one has been equipped with a coupler and a SPDT switch, which periodically changes transmitting antenna into a receiving one.

All devices are based on a universal radiocommunication equipment. The entire system functionality is provided by software installed on computers.

V. RADIO PARAMETERS

Analysing the bandwidth VHF / UHF (Very High Frequen-cy / Ultra High Frequency) among resources available for civil use, it was decided to build system using DS-CDMA technology (Direct Sequence Code Division Multiple Access) in the band 430 MHz, with the following parameters:

- Carrier frequency 431.5 MHz system,
- bandwidth of the transmission channel - 1 MHz (4 MHz)
- Sampling frequency baseband signal - 4 MHz (16 MHz),
- the location information transmission rate - 1 kb / s,
- QPSK modulation (Quadrature Phase Shift Keying).

VI. TESTS AND RESULTS

At the moment there were 4 large test sessions carried out in real conditions. Two sessions were carried out in 2010, in April and October. Then next two in 2011, in June and October. The last test was performed on board a survey ship of the Polish Navy.

The AEGIR system has been tested three times in the Bay of Gdańsk and once - along the coast line of the Baltic Sea. During field tests a position from a satellite navigation system was recorded with the use of a Javad Alpha receiver, which enables simultaneous reception from both American (GPS) and Russian (GLONASS) systems. During the test performed on board of the Navy vessel, two geodetic DGPS receivers have been used (Leica VIVA GS15). They have been installed along the axis of the vessel and the AEGIR antenna has been placed between them.

The effects of our tests performed in October 2010 are illustrated by the visualization shown in Fig. 6, created with use of Google Earth software. The dotted line represents the path of positions received from the satellite systems GPS/GLONASS, and the dots represent the calculated positions of the ground-based system.

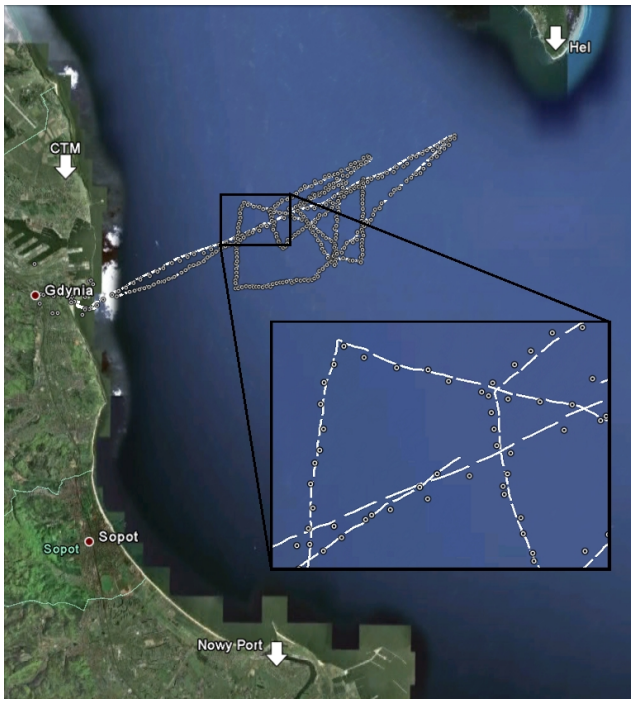


Fig. 6. Deployment of ground stations (arrows) and a path of GPS and GLONASS positions (dotted line) and readings of autonomous AEGIR system (dots)

Analyzing the visualization shown in Fig. 6 it can be observed how accurate the route travelled by the vessel was reconstructed by points calculated by the autonomous localization system – AEGIR.

After completing the measurements, average error (relative to the position shown by the GPS / GLONASS) was obtained at 46m. Distribution and histogram of all results are illustrated in Fig. 7 and Fig. 8.

 ISBN: 978-1-61208-203-5

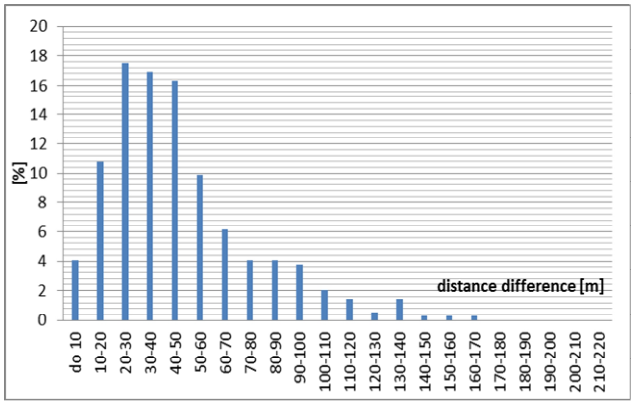


Fig. 7. Histogram of results.

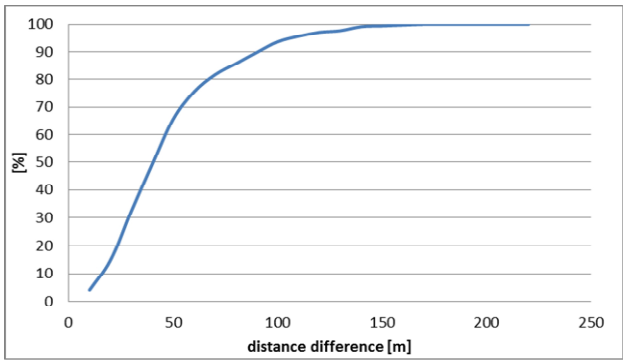


Fig. 8. Distribution of results.

Analyzing the distribution function in Fig. 8 it can be observe, that in 90% of all measured cases the difference distance is bellow 100m.

Test performed in June 2011 was carried out in different configuration. Ground stations were set in a less favorable configuration, along the coastline. Additionally a bandwidth was increased up to 4MHz. It was expected, that because of the unfavorable ground station placement results may worsen comparing to previous ones. However a wider bandwidth would compensate for poor geometry of the system.

Visualization of this configuration is illustrated in Fig. 9.

Fig. 9. Visualisation of AEGIR path when ground station are placed along the shore.

As before an average error has been calculated and obtained at 55m. So as predicted increase of bandwidth nearly compensated a poor system geometry.

Distribution and histogram of obtained results are illustrated in Fig. 10 and Fig. 11.

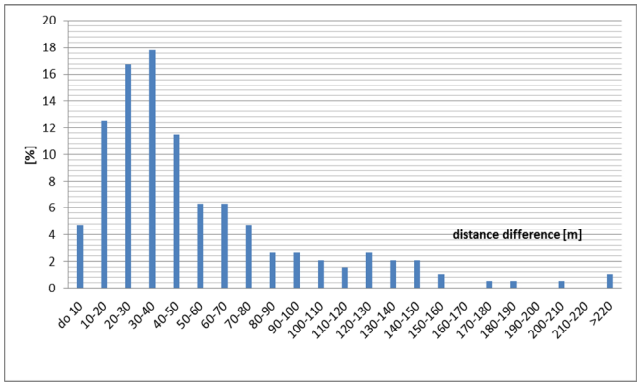


Fig. 10. Histogram of results.

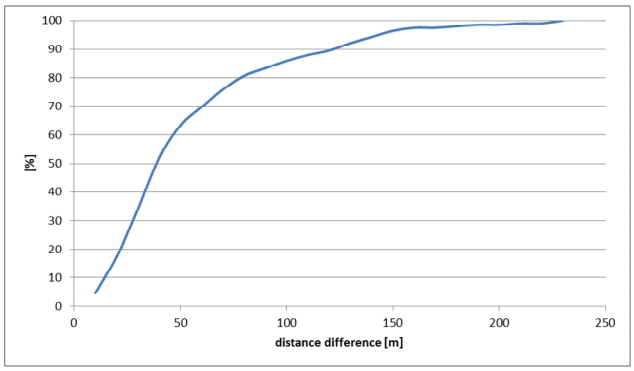


Fig. 11. Distribution of results.

The last test has been carried out in October 2011 with the support of the Hydrographic Office of Polish Navy. The system was tested again in the Bay of Gdansk with the same radio parameters as before. Visualization of travelled route illustrates Fig. 12.

 ISBN: 978-1-61208-203-5

Fig. 12. A path of GNSS positions and readings of autonomous AEGIR system.

The main objective of the tests was the comparison of the results obtained with the use of bandwidth 1MHz and 4MHz. In addition, the correctness of the system was verified in case of location placed outside the geometry of the system (the triangle designated by the reference ground stations). It was expected the coordinates estimated outside the area of good geometry will worsen (outside of the triangle). Figure 13 presents shift of the estimated position outside of the mentioned area.

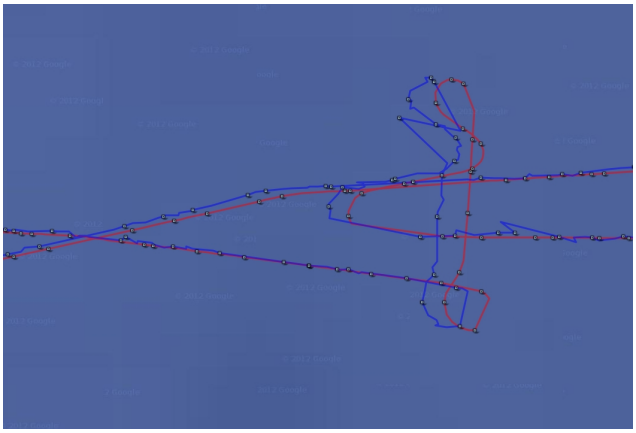


Fig. 13. A path of GNSS positions and readings of autonomous AEGIR system outside of the good geometry of the system

As before an average error has been calculated and obtained at 30m. It should be emphasized that this result also contains the results which have been measured outside the good geometry of the system (outside the triangle).

Distribution of the results is illustrated in Fig. 14.

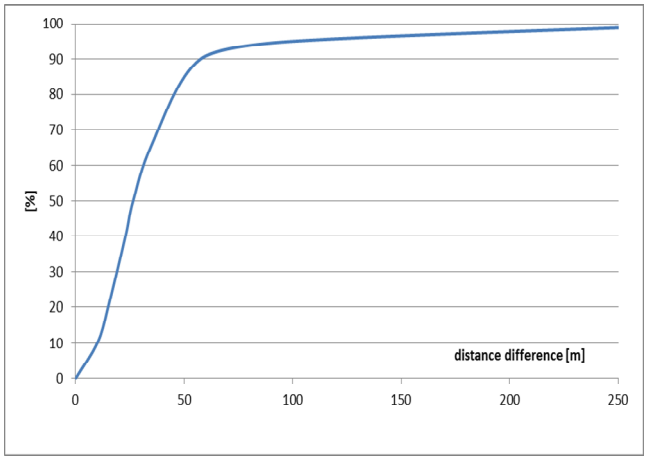


Fig. 14. Distribution of results.

Analyzing the distribution function in Fig. 14 it can be observed, that in 90% of all measured cases the difference distance is bellow 60m.

VII. CONCLUSION

The AEGIR system is a contribution to the development of ground radiolocation systems. The use of modern technology (developed system is fully digital) guarantees the long term operating of such a system.

An important advantage of the developed system is that in the process of estimating the position of the locator coordinates are determined directly (without the need to assign these coordinates to the so-called line items that are needed in phase systems). Therefore, there is no need for start position locator and counting the excess line items that are repeated periodically in the area of the system, depending on the wavelength of the radio and that can lead to errors in the upright-commercial determination of the geographical coordinates of the locator.

The presented system is fully asynchronous. In case of damage or shutdown of one of the stations, the system is fully functional, the only condition is to receive signals from at least three ground stations.

The AEGIR system has been developed to be very flexible. It allows to use more than three ground-stations. Placing them in areas of known positions, allows to create a grid, which will provide ~~an~~ readings of coordinates independent from satellite systems. Using more than three reference station will also increase precision of calculating position.

The AEGIR works in both kind of ground station configuration (triangle deployment and along coastline as well).

 ISBN: 978-1-61208-203-5

In the course of designing and building the system, the adequacy of the system for water bodies has been focused on. However, the versatility of the proposed solution suggests that the system would work on land as well.

VIII. ACKNOWLEDGMENT

The described research was financed from funds for science in the years 2008 - 2011 as a development project.

REFERENCES

[1] *Vulnerability Assessment of the U.S. Transportation Infrastructure that Relies on GPS*, John A. Volpe National Transportations Systems Center, USA, 2001.

[2] *Specification of the Transmitted LORAN-C Signal*, U.S. Department of Transportation and U.S. Coast Guard, May 1994.

[3] Chan Y.T. and Ho K.C., *"A Simple and Efficient Estimator for Hiperbolic Location"*, IEEE Transactions on signal processing, pp. 1905-1915, vol. 42, no. 8, August 1994,

[4] Foy W.H., *"Position-Location Solutions by Taylor-Series Estimation"*, IEEE Trans. On Aero. And Elec. Systems, vol. AES-12, no. 2, 1976, pp. 187-194.

[5] Fang B.T., *"Simple Solution for Hiperbolic and Related Position Fixes"*, IEEE Trans. On Aero. And Elec. Systems, vol 26, no. 5, 1990, pp. 748-753.

[6] Friedlander, *"A Passive Localization Algorithm and Its Accuracy Analysis"*, IEEE Jour. Of Oceanic Engineer., vol. OE-12, no. 1, 1987, pp. 234-245.

[7] Katulski R., Marczak. A., and Stefański J.; *„Software Radio Technology"* (in polish), Telecommunication review and telecommunication news No. 10/2004, pp. 402-406.

[8] Patent application no. P393181, "*Asynchronous system and method for determining position of persons and/or objects*" (in polish), December 2010.

[9] European patent no. EP11460023.2-2220, *Asynchronous system and method for estimating position of persons and/or objects*, May 2011.

 ISBN: 978-1-61208-203-5

EE-AOC: Energy Efficient Always-On-Connectivity Architecture

Sameh Gobriel, Christian Maciocco, Tsung-Yuan Charlie Tai, and Alexander W. Min
Circuits and Systems Research Lab
Intel Labs, Intel Corporation
{*sameh.gobriel, christian.maciocco, charlie.tai, alexander.w.min*}*@intel.com*

***Abstract*—Mobile platforms, such as laptops, tablets or Ultrabooks, must feature *always-on* network connectivity to make mobile applications (e.g., email, instant messaging, etc.) visible and accessible to/from the Internet. Always-on connectivity for mobile platforms can be achieved conventionally by periodically exchanging keep-alive messages with their counterparts (i.e., servers). However, frequent message exchange wastes energy because it requires the platform to operate in (or transition to) the active state (e.g., S0) instead of remaining in a long, uninterrupted low-power standby mode. To address this problem, we present a novel and secure, yet simple, architecture, called *Energy Efficient Always-On-Connectivity* (*EE-AOC*), that allows mobile devices/applications to be continuously visible and reachable from the network. *EE-AOC* offloads the keep-alive message exchange process to the network device from the host, thus, it does not need to wake up the whole platform. Our experimental results implementing the prototype show that *EE-AOC* achieves always-on connectivity and application reachability at about 85% lower power (i.e. 6.67X gain) than that needed to provide the same functionalities in today's platforms.**

Keywords-Always-on-always-connected; connected standby; energy efficient communications; wireless networks; sleep mode.

I. INTRODUCTION

With the skyrocketing use of mobile services and applications in everyday lives, mobile devices, such as tablets, laptops or Ultrabooks, are expected to provide always-on network connectivity anytime anywhere. Unfortunately, always-on network connectivity may incur significant power consumption on mobile platforms because it requires the devices to stay in (or transition to) the active operating state (e.g., S0) [1]—that would require orders of magnitude more energy than low-power standby state (e.g., S3)—to receive and process keep-alive messages, even when the platform is idle. The transition between the standby and active states is very power intensive, and it can drastically shorten battery life, which cannot be tolerated given today's power-hungry mobile devices. Therefore, there is a clear and urgent need for architecture that achieves always-on connectivity for mobile platforms with high energy efficiency.

A typical example of an always-on application is *push email* [2] delivery, made popular by the RIM/Blackberry service, and is now offered by several web-based email services, including Google's Gmail. Users automatically receive email messages as soon as they arrive at the server rather than explicitly having to periodically poll the server to check for new messages. Mobile devices, in this example smart phones, stay in active state (although the display may be switched off) and connected to the network in order to be able to received pushed emails.

While power management features that leverage low-power platform states—i.e., wake up a platform from low-power states only when necessary—have been studied extensively, existing solutions suffer from practical limitations and may not be suitable for mobile platforms. For example, Wake-on-LAN [3] and its wireless equivalent Wake-on-WLAN (WoWLAN) [4] have been proposed as energy efficient means to wake up a platform from the standby S3 state to the active S0 state. However, they are not widely deployed or suitable for mobile devices because they can operate only on a local area network. Moreover, they require modifications to the infrastructure (e.g., access points) to enable wider, Internet-related usages, making them insufficient to fully realize always-on connectivity for mobile platforms.

In this paper, we present a novel architecture, called Energy Efficient Always-On-Connectivity (*EE-AOC*), which enables the low-power operation of always-on and always-connected usage models for mobile platforms. *EE-AOC* allows mobile platforms to enter a low-power sleep state, e.g., S3, without the risk of losing network connectivity, thereby making them continuously reachable to/from the application servers on the Internet. *EE-AOC* offloads the processing of the protocols required for preserving network connectivity to the wireless network communication device (W-NIC) instead of processing them in the host. *EE-AOC* has the following salient features:

- W-NIC in *EE-AOC* handles protocols necessary for maintaining network connectivity, including link layer key refresh, address resolution protocol (ARP) [5], and the presence protocols for various application servers, in a highly energy efficient manner, consuming slightly more energy than the S3 standby state energy consumption. However, *EE-AOC* does not require the W-NIC to have full network stack processing capabilities, hence reducing device cost and ensuring continuous availability.
- W-NIC in *EE-AOC* has the ability to wake-up the platform to the active state upon reception of a generic

 ISBN: 978-1-61208-203-5

network packet from any Internet device. This extends the capability of WoWLAN, which only supports wake patterns based on a magic packets (e.g., a broadcast frame containing a specific number of repetitions of the target device 48-bit MAC address) transmitted on a local area network [4].

- *EE-AOC* allows a platform in sleep state to be woken up securely by targeted Internet services, being highly robust against Denial-of-Service (DoS) attacks, such as an energy drain attacks caused by constant malicious wakes.

We would like to note that although we focus on TCP-based [6] presence protocols, *EE-AOC* is not restricted to TCP and can be applied to any network protocols.

The main contributions of this paper can be summarized as follows:

- We design an energy efficient architecture for mobile platforms, called *EE-AOC*, that maintains device connectivity and application presence to the Internet by delegating the keep-alive message protocol processing to the W-NIC, while the platform remains in a low-power standby state.
- We prototype *EE-AOC* by modifying the firmware of an off-the-shelf Intel WiFi NIC, to demonstrate its efficiency for AOAC usages, e.g., IMAP [7] based pushed emails, and compute continuum [8] usage models. Our in-depth evaluation results show that *EE-AOC* saves about 85% of the overall platform energy.

The rest of the paper is organized as follows. In the next section, we present a literature review related work. Section III highlights the platform energy consumption to maintain network connectivity. Section IV presents our *EE-AOC* framework and describes its architecture and algorithms for offloading the network connectivity and presence maintenance to the W-NIC. Section V evaluates the proposed *EE-AOC* technology and presents the implementation results. We conclude the paper in Section VI.

II. Related Work

Several mechanisms have been proposed to reduce the energy consumption of networked platforms, and various regulatory organizations worldwide are now mandating improvement in the energy consumption of platforms in standby state [9]. Prior proposals can generally be grouped into three categories: (i) those that reduce the active power consumption of systems [10]–[12], (ii) those that reduce the power consumption of the network infrastructure, routers and switches [13]–[15], and (iii) those that opportunistically put the devices to sleep [16]–[18].

EE-AOC falls into the third category and advocates for localized energy-efficient optimization within the platform to extend the sleeping state while maintaining network connectivity, presence and reachability. *EE-AOC* uses the platform W-NIC device for both wake-up and presence, rather than using a network-wide implementation of a proxy service [19]. As mentioned previously, *EE-AOC* is not restricted to a magic packet as defined in WoWLAN, and it supports any wake patterns securely negotiated between the client and the server. In [16], network presence, reachability and application support are realized through the use of an offload processor and application program modifications. In this paper, we show that *EE-AOC* achieves this goal at a very low energy level, without requiring additional hardware. *EE-AOC* saves significant platform energy because it interrupts the host platform only when application support is required.

III. Problem Statement

In this section, we analyze and quantify the potential platform power implications when the platform low-power state is continuously interrupted to exchange keep-alive messages.

A. Mobile Platform Battery Lifetime

Current platform power management policies guide the platform to enter a low-power sleep state (i.e., Sx) when the platform becomes idle for a pre-defined period of time. In general, the longer the system stays in sleep state uninterrupted, the higher the energy gain is, because more peripheral devices and system components can enter a deeper low-power state for a longer period of time.

Recent work has shown that smart timing approaches for Operating Systems (OS) can be used to increase the quiet (idle) time for the OS by skipping timer tick interrupts when the platform is idle or by adaptively changing the rate of timer interrupts (e.g., [20]). This forms the basis for *"tickless OS"*, in which the OS is moving away from scheduling periodic clock interrupts every few milliseconds to a more event-driven approach [21], [22], in which the OS is woken-up to process an event being posted by the applications or by the hardware on demand.

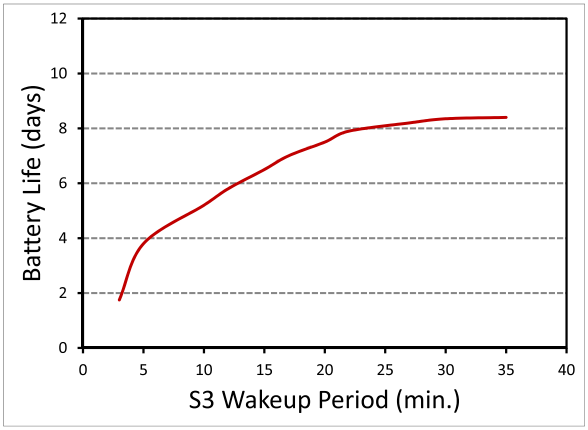


Figure 1. Platform Battery Life vs. S3 Time

 ISBN: 978-1-61208-203-5

Figure 1 illustrates the relationship between typical platform battery life (in days) and sleep time, i.e., the period of time the platform resides in the sleep state (i.e., S3) before transitioning to the active state (i.e., S0). The concavity of the curve indicates that the battery life is sensitive to the sleep time (or the wakeup frequency). It shows that the battery depletes much faster with increasing wakeup frequency and becomes less sensitive to wakeup frequency at longer sleep times (e.g., > 25 min). The concavity of the curve can indeed be attributed to the considerable energy overhead caused by platform transitions from the sleep state to the active state.

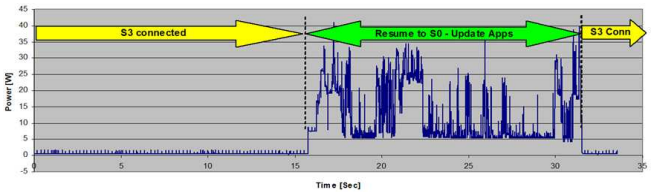


Figure 2. Power Waveform for S3-S0 Transition

Figure 2 shows the power measurement of a laptop when it wakes up from S3 to S0. It indicates that the energy consumed in waking up (i.e, the area under the curve from 16s to 32s) is much higher than the energy consumed when the platform is in sleep state (which is the flat line near 0). The time duration in which the platform stays in the active state before going back to the low-power state depends on various factors, such as the OS scheduling policy, the number of applications running, network connection time, the amount of data exchanged, and network traffic characteristics. Therefore, it is important to minimize the energy overhead due to state transitions to achieve energy-efficient always-on network connectivity.

B. AOAC and Wakeup Frequency

As we mentioned earlier, Always-On-Always-Connected (AOAC) capabilities are crucial for mobile platforms to receive "pushed" data (e.g., an IM message or a tweet update, etc.) from an Internet server in real-time. AOAC is also important for several usage models, in which remote devices need to be discovered and paired together, and thus must be reachable across the Internet (e.g., when a user wants to use his smartphone at a hotel to watch a movie he downloaded on his laptop left at home.)

As a result, connectivity between client platforms and Internet servers must be maintained, which is typically accomplished by exchanging "keep-alive" messages at fixed time intervals. When a connection is established between an AOAC application client (e.g., IM client) and AOAC application servers (e.g., IM server) the connection is kept alive until it expires after a timeout of "T" seconds. If no data is exchanged over this connection for longer than T seconds, the connection is dropped, and the client becomes unreachable by the server (i.e., the server marks the client as offline). Once the connection is dropped, new data (e.g., new IM message sent to the client) will no longer be pushed to the client. However, exchanging keep-alive packets is an energy consuming task because it either requires the client to be in the active state or to transition from a low-power state to the active state.

A key parameter that determines the lifetime of an AOAC device is the frequency of the keep-alive messages. The maximum value of T should be the minimum of T_S and T_N, where T_S is the timeout of the application server and T_N is the minimum timeout of any communication equipment (e.g., Network Address Translation "NAT" box [23]) along the route from the client to the server. These timeouts are in place mainly because each connection has a state (e.g., a state may include source and destination addresses and port numbers) that must be maintained. Due to limited resources and/or improved responsiveness (e.g., faster offline indication) the storage time of the connection state cannot be indefinite and will be dropped after a given timeout. Therefore, AOAC devices must exchange keep-alive messages in a timely manner in order to maintain network connectivity.

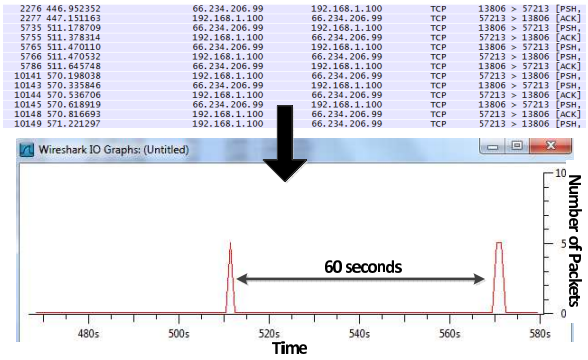


Figure 3. Skype IM Application Keep-Alive Message Period

However, it is practically infeasible to accurately predict connection timeout values (i.e., T) at the design stage because of their inherent variability and the unpredictability of the network path between client and server. For example, from the application server side, a typical TCP default session timeout is 2 hours. On the other hand, from the client side, an IM offline indication is usually in the range of 3-5 minutes. Figure 3 shows the network packet trace of a Skype application, in which the client exchanges keep-alive messages with the server approximately once every 60 seconds. Moreover, the network timeout has a wide range of values. Default factory settings of an NAT device timeout can typically range from as short as 15 seconds to as long as one hour. Therefore, AOAC devices need to use a shorter timeout (in the range of tens of seconds) for keep-alive period, i.e., T, which may have a significant power impact as shown in Figure 1. In the next section, we elaborate on how *EE-AOC* overcomes these practical challenges.

 ISBN: 978-1-61208-203-5

IV. EE-AOC TECHNOLOGY

EE-AOC achieves the functionality of exchanging keep-alives with the Internet application server, and hence, maintains the connectivity, presence and reachability of an AOAC device to the network at very low power. This is achieved by offloading a series of "keep-alive" packets to the wireless network interface card (W-NIC), which impersonates the platform in a low-power state to other hosts and servers on the network.

Furthermore, *EE-AOC* allows the NIC to maintain a TCP keep-alive session to the server, modifying only the NIC firmware, without changing the NIC hardware. More importantly, by using TCP at the server side, the network infrastructure pipes (e.g., network switches, network wireless access points, etc.) do not require changes and are not even aware that the platform is not running in the active mode.

A. Always-On-Always-Connected System Architecture

In an AOAC system, clients maintain connectivity with the Internet server and keep the network pipe open with keep-alive messages to the server. For example, when there is application data to be pushed to the client, e.g., an IM message or an email, the server will use the established client-server session. Like most Internet traffic, the established session is based on TCP protocol. Consequently, in order for the client to receive live updates from a set of AOAC applications, it has to maintain an alive session with the corresponding server for each supported AOAC application. Obviously, the overhead of connection maintenance increases with the number of established connections.

When the client enters a low-power state between the transmission of keep-alive messages, then from an energy-efficient system design standpoint, it is crucial to maximize the time the client spends in a low power state. Having multiple ongoing AOAC connections, each with its own periodicity, leads to unaligned timing in sending the keep-alive messages, and as a result, the client is forced to exit the low-power state more often.

To reduce unaligned keep-alive message periods, an AOAC system uses a push server, as shown in Figure 4. A typical example of such architecture is Apple Push Notification Service (APNS) [24]. In this case, the client maintains the connectivity and an alive session with only one Internet push server. The server aggregates and proxies data updates for other application servers. As shown in Figure 4, when an update is available for the user, the application server sends a notification to the push server, which then pushes this notification to the client. Upon reception of the notification, the client either (i) exits the low-power state to an active state, so that it can receive the data update from the push server or (ii) can establish a new connection to the application server to retrieve the available update.

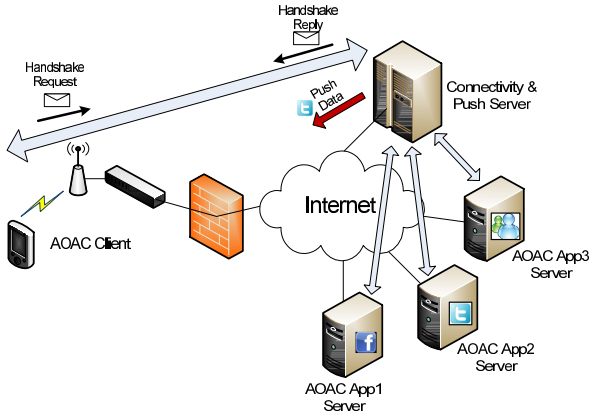


Figure 4. AOAC System Architecture

B. EE-AOC and Network Infrastructure

The client is connected to the Internet through a network infrastructure (i.e., routers NAT boxes, proxy servers, etc.), and each component of this infrastructure will keep a state for the client as long as the connection is maintained. Soon after the connection terminates, or if it is not maintained, the network infrastructure will drop the client state to save resources. Then, the client will need to re-initiate its state in the network for new connections.

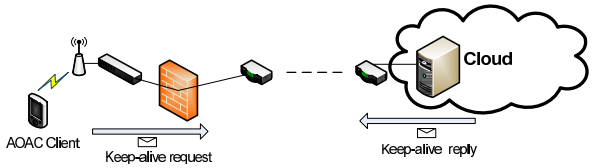


Figure 5. Keep-Alives Handshake through Network Infrastructure

An example of such a network state is firewall flow state that does not block traffic as long as it is part of an existing flow. As shown in Figure 5, when the client initiates the connection, the firewall will check the connection state to decide whether or not it is permissible or not. If the connection is legitimate, then the firewall will keep the traffic belonging to this flow going through and will not block the traffic. When the firewall detects that the traffic has stopped, e.g., with no TCP FIN (i.e., flow end message) being detected, it will drop the state of this flow after a short timeout, unless new packets belonging to this flow are identified. In this case, the client must reset the cycle and reinitiate the request to connect to the server. On the other hand, most firewalls will block the traffic if the flow state is dropped and new data or a new connection request is sent form the Internet cloud server to the client because most firewalls do not allow connections to be initiated from outside.

Therefore, protocols for keep-alive messages must be able to traverse the network infrastructure with default settings

 ISBN: 978-1-61208-203-5

(i.e., no special ports to be opened or no special changes to the security policy). Furthermore, the keep-alive handshake process has to be initiated by the client because a server-initiated handshake will not work in the absence of an alive session. In section IV-D, we highlight how our *EE-AOC* technology achieves the above-mentioned properties.

C. EE-AOC and Secure Wake

AOAC clients are always connected to the network, waiting for data updates to be pushed by the push server. This always-on connectivity, hence longer exposure to the Internet, makes mobile clients vulnerable to attacks. An attacker can launch a Denial-of-Service (DoS) attack on AOAC clients by pushing frequent wake-up messages to deplete their batteries faster (see Figure 1), thereby rendering them unreachable to legitimate users.

Another possible attack scenario is that an attacker can impersonate the client by sending fake keep-alive messages to the server on behalf of the legitimate AOAC client. As a result, the server can be misled to believe that the legitimate client is still available and connected to the network. This may cause a data integrity violation because the client's presence can no longer be trusted. Even worse, it can cause data loss if the server forwards the client's private data to the attacker.

Therefore, the keep-alive message exchange must be designed to provide integrity and authenticity of the data exchanged between the client and server. In the next section, we discuss the secure design and overall operations of *EE-AOC*.

D. EE-AOC Software Architecture

EE-AOC enables end-users' software applications/services to be connected to the network application server, while the platform remains in a low-power standby mode. As we mentioned earlier, these applications/services can maintain connectivity by periodically sending keep-alive messages to the application servers. To achieve this goal in an energy-efficient manner, *EE-AOC* defines an interface and network device functionality and capability as we describe next.

AOAC applications intending to maintain connectivity and presence to the network pre-build a list of keep-alive messages for the next few minutes, using each application's proprietary protocol, appropriate sequence number, and periodicity information (if required). Then they are secured with the application key/tokens, and handed over to the communication device along with the appropriate information about each message (e.g., required periodicity to maintain presence, where should it be sent (to which address), etc.) before the platform transitions into the standby state.

Upon transitioning to the low-power state, the communication device performs the following three operations: (i) it recovers the keep-alive messages, (ii) it orders them chronologically, and (iii) it sends these pre-build keep- alive messages to their destinations network at the appropriate time in order to maintain its presence to the application servers. The outline of the *EE-AOC* architecture and the offload algorithm is highlighted in Figure 6.

It should be noted that the idea of offloading TCP has been proposed before. However, full TCP offload is very complex and consumes a considerable amount of scarce resources (e.g., processing, memory, etc.) from the network interface card, to the extent that can hardly be supported by client network card vendors. Although *EE-AOC* defines an offload framework for the NIC, it is simple and requires no modifications to hardware or additional resources from the network card, as will be evident later.

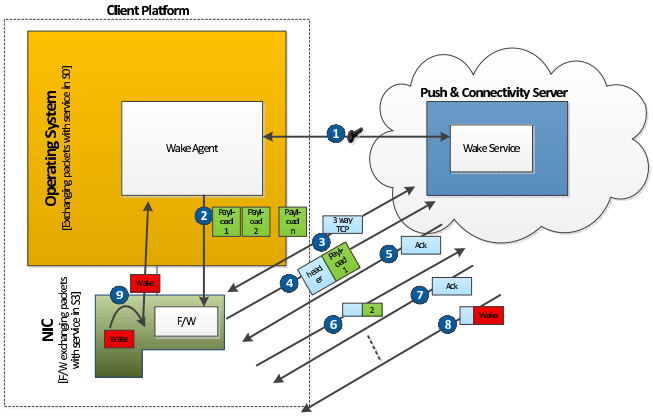


Figure 6. EE-AOC Offload Architecture and Operation

As shown in Figure 6, the offload operations can be divided into two phases. The first phase is to prepare a list of keep-alive messages, and this is executed while the platform is in the active mode; this phase runs inside the wake agent as part of the OS or as an application. The second phase is to exchange the packets with the Internet server in a timely manner to keep the connection alive. This will run inside the NIC firmware when the platform is in standby or a low-power state, while the NIC is in the operating state. We elaborate on the detailed steps of these two phases as follows:

In the platform active state before going to standby:

1) The wake agent registers the platform with the Internet server.
2) The wake agent exchanges keys with the server and establishes a shared private key with the server. [Arrow 1 in Figure 6]
3) The wake agent prepares a train of payloads (multiple payloads) that are encrypted with the shared private key.
4) The wake agent formats the train of packets as http post messages (destination port 80) on top of TCP. Since HTTP/TCP connections are legitimate inside out connections, no firewall will block the connection.

 ISBN: 978-1-61208-203-5

5) The wake agent adds TCP-SYN packet to the head of the packet train, followed by an ACK packet. (needed for the 3-way TCP session initiation handshake)
6) The wake agent transfers the train of packets to the NIC driver. [Arrow 2 in Figure 6]
7) The wake agent downloads the wake filter to the NIC. This is a pattern of a general packet that once the NIC receives and matches, it should wake the platform up.
8) When the driver detects the OS event that indicates that the platform is transitioning to standby it loads the NIC firmware image with the train of packets.

In the standby/low power platform state:

1) The NIC blindly transmits the first packet it got from the train of packets. As mentioned, the first packet is TCP-SYN
2) The server replies to the TCP-SYN with a SYN-ACK message that has its own sequence number that the NIC firmware will extract from the packet and send the Ack message (2nd packet in the train of packets) accordingly. [Arrow 3 in Figure 6]
3) Once the TCP flow is initiated with the server, the NIC sends the http-formated keep-alive packets downloaded from the wake agent.[Arrow 4 in Figure 6]
4) The server decrypts the keep-alive packet, and if the decrypted packet is correct, then an ack message is sent to the client. [Arrow 5 in Figure 6]
5) The NIC keeps a retransmission window of 1 packet and waits for an ack of the packet from the server. If the ack is not received, then the packet is retransmitted for a maximum number of retries. Keeping a window of 1 packet eliminates the need for a complete TCP stack offload because the NIC firmware does not need to handle the sizing of the TCP congestion window or any optional flags in the TCP header.
6) If the maximum number of retries (or any other exception, e.g., platform is out of network coverage) the NIC wakes up the platform to S0 to handle this case.
7) The NIC goes on the train of packet payloads it has in its firmware, sending them one by one to the server. This is done at a pre-determined frequency, typically once every minute. [Arrow 6 in Figure 6]
8) The NIC keeps sending the keep-alive packets until the whole list of packets has been exhausted, and in this case, the NIC wakes up the platform to active state to restart, where the wake agent re-exchanges keys and prepare a new list of packets.
9) If while in standby there is new data to be sent from the server to the client (outside in data), the server simply sends the encrypted wake message as the payload to a TCP message, formatted as an http reply. [Arrow 8 in Figure 6]
10) The NIC decrypts the received packet, and if the wake packet content matches the pre-defined wake filter, then the NIC wakes the platform up to active state.
11) In active state, the application connects to the server to download the new data received, and when this is finished, the platform goes back to standby and the keep alive cycle restarts.

V. Performance Evaluation

We implemented the system described in Figure 4. Specifically, we used the Intel WiFi 5350 [25] NIC firmware to offload the secure keep-alive messages. We had a wake service running on the Internet server to aggregate and push updates to the client when it was in a low-power state. We used two exemplary applications, i.e., Facebook and email, to demonstrate the benefits of *EE-AOC*. We implemented the wake server to periodically monitor the Facebook account of the client and push updates posted on his Facebook wall towards the client. Similarly for an email application, we used push Internet Message Access Protocol (p-IMAP) [26] to push new emails sent to the client as soon as they arrived at the mail server. In our evaluation, we analyze the power consumption and battery life of the platform when using *EE-AOC*.

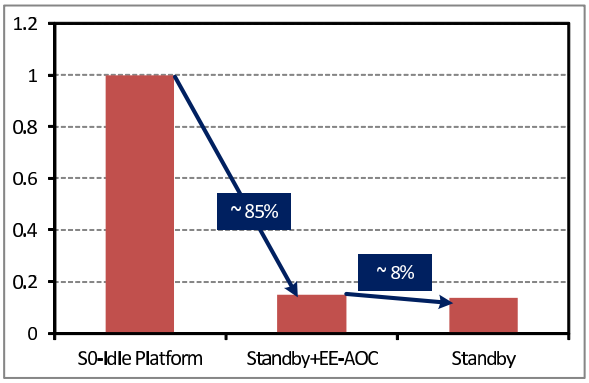


Figure 7. Normalized Power Consumption Comparison

Figure 7 shows the power consumption of a laptop in three operating states. (1) *Fully-on*, that is when the platform is switched on and stays in the active state (i.e., S0), but no active workload is running. (2) *Standby with* EE-AOC *enabled*, that is when the platform is in standby, but*EE-AOC* is implemented in the NIC, which entails that the communication device is turned on and is exchanging the keep-alive messages with the service to maintain connectivity and presence. (3) *Standby*, that is when the platform is in sleep state with no network connectivity. The power consumption in Figure 7 is normalized to the *fully-on* state. This is because we believe that although the absolute power consumption in each of these states will decrease from one generation to the next, and will be different based on the platform itself, the relative power consumption among these states will remain relatively unchanged.

 ISBN: 978-1-61208-203-5

As shown in Figure 7, there is a significant difference in power consumption between the *fully-on* state and the *standby* state. In standby, (a.k.a. suspend to RAM), most of the platform components are turned off and the operating system and application states are saved to the RAM which remains powered. The platform in standby consumes only about 10 % of the power consumed in the *fully-on* state, confirming that standby is a very efficient power saving mode for the client. On the other hand, *EE-AOC* requires additional power for the network communication device, which remains connected to the network, sending and receiving data packets. However, The NIC power consumption is typically very small when compared to the whole platform. Thus, the power consumption of *EE-AOC* is slightly higher ($\sim$8 %) than that of standby mode power consumption, yet significantly lower than lighting up the whole platform.

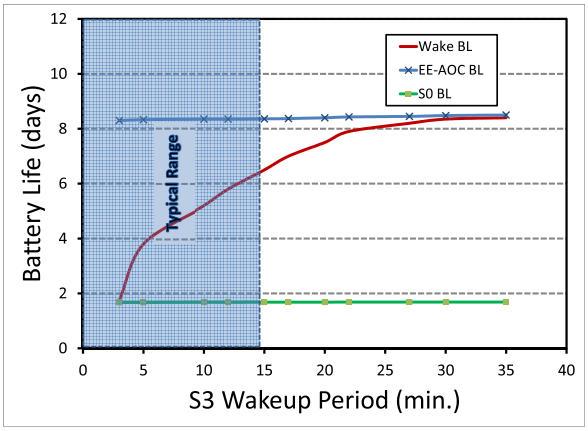


Figure 8. Battery Life vs. Sleep Time with EE-AOC

Figure 8 shows the battery life (in days) of an AOAC client with respect to the frequency of exchanging the keep-alive messages for the following three cases: (1) when the platform is in the fully-on mode and exchanges the keep-alive messages (denoted as S0 BL), (2) when the platform enters the standby state between keep-alive messages (which is the result discussed in Figure 1) and wakes up to S0 when it is time to send a new keep-alive message (denoted as Wake BL), and (3) when the keep-alive messages are offloaded to the NIC with *EE-AOC* (denoted as *EE-AOC* BL).

Figure 8 shows that when the platform stays in the fully-on mode, the battery life of the platform is very short and will only last for less than two days. In this case, the curve is almost constant and the battery life depends less on the periodicity of the keep-alive messages because the energy consumption for transmitting and receiving a packet is negligible when compared the power consumption of the whole platform. However, when the platform enters the standby state between the keep-alive packets, energy consumption is dominated by the overhead of entering into and exiting from the sleep state, leading to the concave-shaped curve of battery life. Unfortunately, without *EE-AOC*, most applications/networks require the keep-alives to be exchanged in the order of every tens of seconds, which leads to a poor platform battery life.

On the other hand, when *EE-AOC* is used, the network communication device exchanges the keep-alive messages on behalf of the platform. Most of the platform components will stay in a low-power mode, while the NIC is turned on to serve the keep-alive messages. The battery in such a case lasts longer than 8 days, which is more than a 4X increase. Moreover, similar to the fully-on case, the curve is almost flat because the energy of sending and receiving packets is small relative to the NIC power. This indicates that the battery life with *EE-AOC* is almost constant, irrespective of the frequency of the packets exchanged.

VI. Conclusion and Future Work

Conventionally, platform power management policies are designed to guide individual platform components or the whole platform into low-power sleep states when the platform becomes “idle”, with no active workloads. Individual power management techniques differ in how deep the sleep state is, the algorithms used to enter and exit the sleep states, and the optimizations to extend these sleep states as long as possible. In this paper, we proposed a novel, yet simple, architecture, called *EE-AOC*, that achieves energy-efficient always-on network connectivity for mobile platforms. *EE-AOC* is very different from existing solutions in a sense that it does not require hardware modifications, while it provides the core functionalities for Always-On-Always-Connected (AOAC) usage models and application visibility to mobile platforms, such as Ultrabooks, laptops and tablets.

As future work we’ll explore how the communication device can help optimize the overall platform power when the platform is in the active state (i.e., S0) either idle with no active applications running or lightly loaded with various tasks. As the platform’s energy in S0 is optimized with the new generation of platforms, any help provided by peripheral devices to optimize its energy consumption will benefit the end-users with longer battery life

References

[1] “ACPI advanced configuration and power interface specifications rev 4.0,” http://www.ecma-international.org/, November 2009, [Online; accessed 08-June-2012].

[2] Z. Duan, K. Gopalan, and Y. Dong, “Push vs. pull: Implications of protocol design on controlling unwanted traffic,” in *SRUTI*, 2005, pp. 25–30.

[3] “Wake-On-Lan,” http://en.wikipedia.org/wiki/Wake-on-LAN, [Online; accessed 08-June-2012].

[4] N. Mishra, K. Chebrolu, B. Raman, and A. Pathak, “Wake-on-wlan,” in *international conference on World Wide Web (WWW)*, 2006, pp. 761–769.

 ISBN: 978-1-61208-203-5

[5] D. C. Plummer, “An Ethernet Address Resolution Protocol,” http://tools.ietf.org/html/rfc826, November 1982, [Online; accessed 08-June-2012].

[6] “Transmission control protocol, protocol specification,” http://www.ietf.org/rfc/rfc793.txt, September 1981, [Online; accessed 08-June-2012].

[7] M. Crispin, “Internet message access protocol,” http://tools.ietf.org/html/rfc3501, March 2003, [Online; accessed 08-June-2012].

[8] J. Henrys, “The compute continuum: A wave of connected devices,” http://www.internetviz-newsletters.com/eletra/mod_print_view.cfm?this_id=1940237&u=intel&show_issue_date=F&issue_id=000481299&lid=b11&uid=0, [Online; accessed 08-June-2012].

[9] “ENERGY STAR: program requirements for computers,” http://www.energystar.gov/ia/partners/prod_development/revisions/downloads/computer/Version5.0_Computer_Spec.pdf, [Online; accessed 08-June-2012].

[10] Y. Agarwal, T. Pering, R. Want, and R. Gupta, “SwitchR: Reducing System Power Consumption in Multi-Clients Multi Radio Environment,” in *IEEE International Symposium on Wearable Computers (ISWC)*, 2008, pp. 99–102.

[11] J. Flinn and M. Satyanarayanan, “Managing Battery Lifetime with Energy Aware Adaptation,” in *ACM Transactions on Computer Systems*, vol. 22, 2004, pp. 137–179.

[12] X. Li, R. Gupta, S. Adve, and Y. Zhou, “Cross Component Energy Management: Joint Adaptation of Processor and Memory,” in *ACM Transactions on Architure and Code Optimization*, vol. 4, no. 14, 2007.

[13] C. Gunaratne, K. Christensen, and B. Nordman, “Managing Energy Consumption Costs in Desktop PCs and LAN Switching with Proxying, Split TCP Connections and Scaling of Link Speed,” in *International Journal of Network Management*, 2005, pp. 297–310.

[14] M. Gupta and S. Singh, “Greening of the Internet,” in *ACM Sigcomm*, 2003, pp. 19–26.

[15] S. Nedevschi, L. Popa, G. Iannaccone, S. Ratnasamy, and D. Wetherall, “Reducing Network Energy Consumption via Sleeping and Rate Adaptation,” in *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2008, pp. 323–336.

[16] Y. Agarawal, S. Hodges, R. Chandra, J. Scott, P. Bahl, and R. Gupta, “Somniloquy: Augmenting Network Interfaces to Reduce PC Energy Usage,” in *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2009, pp. 365–380.

[17] E. Shih, P. Bahl, and M. Sinclair, “Wake on wireless: An event driven energy saving strategy for battery operated devices,” in *IEEE International Conference on Mobile Computing and Networking (MobiCom)*, 2002, pp. 160–171.

[18] J. Sorber, N. Banerjee, M. Corner, and S. Rollins, “Turducken: Hierarchical Power Management for Mobile Devices,” in *IEEE International Conference on Mobile Systems, Applications and Services (MobiSys)*, 2005, pp. 261–274.

[19] “proxZZZy for sleeping hosts,” http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-393.pdf, February 2010, [Online; accessed 08-June-2012].

[20] C. Olsen and C. Narayanaswami, “PowerNap: An Efficient Power Management Scheme for Mobile Devices,” in *IEEE Transactions on Mobile Computing*, 2006, pp. 816–828.

[21] V. Yodaiken and M. Barabanov, “A Real-Time Linux,” in *Linux Journal*, vol. 34, 1997.

[22] T. Gleixner and I. Molnar, “Dynamic Ticks,” http://lwn.net/Articles/202319/, October 2006, [Online; accessed 08-June-2012].

[23] K. Egevang, C. Communications, and P. Francis, “The IP Network Address Translator (NAT),” http://www.ietf.org/rfc/rfc1631.txt, [Online; accessed 08-June-2012].

[24] “Apple push notification service,” https://developer.apple.com/library/mac/documentation/NetworkingInternet/Conceptual/RemoteNotificationsPG/ApplePushService/ApplePushService.html, [Online; accessed 08-June-2012].

[25] “Intel WiFi Link 5100 Series specifications,” www.intel.com, 2008.

[26] S. H. Maes, C. Kuang, R. Lima, R. Cromwell, E. Chiu, J. Day, R. Ahad, W.-H. Jeong, and G. Rosell, “Push extensions to the imap protocol (P-IMAP),” http://tools.ietf.org/html/draft-maes-lemonade-p-imap-12, March 2006, [Online; accessed 08-June-2012].

 ISBN: 978-1-61208-203-5

TAO: A Time-Aware Opportunistic Routing Protocol for Service Invocation in Intermittently Connected Networks

Ali Makke, Nicolas Le Sommer and Yves Mahéo
{Ali.Makke,Nicolas.Le-Sommer,Yves.Maheo}@univ-ubs.fr
IRISA, Université de Bretagne-Sud, France.

***Abstract*—Handheld devices owned by nomadic people can form intermittently connected mobile ad hoc networks spontaneously. Such networks appear as an attractive solution for service providers, such as local authorities, in order to extend a pre-existing infrastructure-based network composed of several infostations so as to provide nomadic people with application services in a large scale area (e.g., a city). In such hybrid networks, intermittent connections are prevalent, and end-to-end paths between clients and providers cannot be maintained all the time. Service provisioning thus remains a challenging problem today in these networks. In this paper, we propose a new time-aware opportunistic routing protocol called TAO. TAO is designed for service invocation in intermittently connected hybrid networks. This protocol makes it possible to select the best next message forwarder(s) among a set of neighbor nodes based on the dates of contacts of these nodes with infostations, and tends to implicitly estimate the distance separating these mobile nodes and the infostations. This paper gives a detailed description of this protocol supported with some simulation results.**

Keywords-Service invocation; Opportunistic Networking; Disconnected Mobile Ad hoc Networks.

I. INTRODUCTION

With the increasing proliferation of handheld devices equipped with a wireless interface, such as smart-phones or internet tablets, new kinds of applications, services or networks relying on spontaneous communication, interaction and collaboration can be considered. Such devices, which can form mobile ad hoc networks spontaneously, can be exploited in order to extend networks composed of some sparsely distributed infostations across a large area, (e.g., a city), and so to create hybrid networks in which infostations can act as service providers and mobile devices as service clients. This kind of networks, illustrated in Figure 1, is said hybrid because they are formed of a fixed part (the set of infostations connected together via a wired infrastructure) and a mobile part (the ad hoc network of mobiles nodes). Hybrid networks could appear as an opportunity for service providers, such as local authorities, to provide nomadic people with new ubiquitous services, without resorting to any expensive infrastructure, such as those provided by Global System for Mobile Communications (GSM) operator companies.

The fixed part of these hybrid networks can obviously present various topologies. For instance, the services can be provided by dedicated servers that can be accessed by the mobile devices through the infostations, which act as gateways. We focus in this paper on message routing in the mobile part of the hybrid network. Consequently, we will assume, without loss of generality, that a service is provided by an infostation, directly or via another infostation.

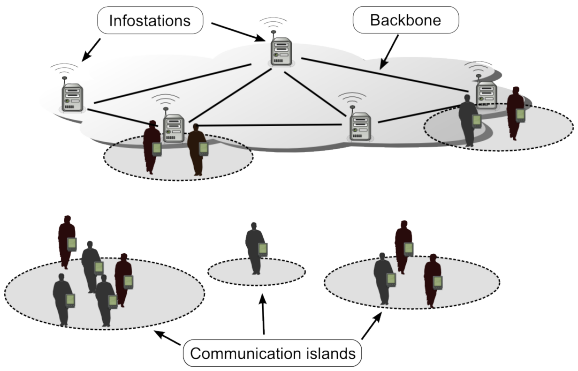


Figure 1. Example of an intermittently connected hybrid network.

Both the free movements of people and the short communication range of Wi-Fi wireless interfaces accompanied with the radio interferences induce some frequent and unpredictable disruptions in the communication links. These disruptions entail the creation of disconnected communication-islands formed near or far from the infostations embedded in the physical environment (see Figure 1). Moreover, the volatility of the devices, which are frequently switched off due to their limited power budget, increases the number of changes in the network topology. In these networks, which we will qualify as intermittently connected hybrid networks (ICHN), it is difficult, and most often even impossible, to maintain an end-to-end path between two devices thanks to traditional routing protocols designed for wired networks or thanks to dynamic routing protocols such as Ad hoc On Demand Distance Vector (AODV) or Optimized Link State Routing Protocol (OLSR).

To cope with these issues, routing protocols devised for networks that suffer from frequent and unpredictable disruptions, such as delay tolerant and opportunistic networks, implement the "store, carry and forward" general principle. When two devices cannot communicate directly because they are not in the transmission range of each other, these protocols make it possible to exploit other mobile nodes as intermediate relays

 ISBN: 978-1-61208-203-5

that can carry a copy of a message when they move and forward it afterwards to other nodes so that it eventually reaches its destination. The provision of application services with this kind of opportunistic and asynchronous communications has been addressed so far only by a few number of research works [1], [2], [3], [4]. Mahéo and Said [1], and Conti and Kumar [3] focus on service provisioning in opportunistic networks composed solely of mobile nodes. In [1], the authors propose content-based service discovery and invocation solutions in order to exploit the redundancy of the services offered by the mobile devices that can move freely (i.e., no assumptions are made regarding the mobility of the devices). Conti and Kumar [3] target networks relying on social interactions between mobile nodes that act as both clients and providers of services. Due to the volatility and the limited resources of the mobile devices, the number of relevant services that can be offered by these devices are limited in comparison to those that could be offered in hybrid networks. Unlike [1] and [3], in Le Sommer et al. [2] the services are provided by fixed infostations in limited geographical areas. In [2], mobile devices and infostations are aware of their own location. Mobile devices can invoke remote infostations thanks to an opportunistic and location-aware forwarding protocol. In contrast with the environments we consider, in [2], the infostations were not connected together. We believe that the design of routing protocols suited for service discovery and service invocation in ICHN should take the specificity of ICHN into account, namely the interconnection of the infostations, in particular when defining heuristics or functions that allow to select among a set of neighbor nodes the node(s) that are considered as the best relay(s) to reach one of the infostations.

In this paper, we propose TAO, a new opportunistic and time-aware routing protocol devoted to service invocation in ICHN. TAO adopts a temporal heuristic to perform efficient message forwarding decisions. The basis of our approach is to take into account the disconnection dates between the fixed infostations and the mobile nodes. Assuming a homogeneous speed of the mobile nodes, a node implicitly estimates the distance from its neighbors to the target infostation and is able to select the best carriers among them. Our proposal applies to situations in which routing decisions cannot be taken according to geolocation information.

The rest of the paper is structured as follows. Section II describes protocol TAO, and Section III the evaluation of this protocol. Related works are discussed in Section IV. Section V presents our conclusions and gives some perspectives.

II. The TAO Protocol

TAO aims at supporting the invocation of software services in ICHN such as those formed by fixed infostations and portable devices used by nomadic people. The description of the discovery phase that should occurs before the invocation is not in the scope of this paper. The remainder of this section details how service invocation requests and responses are forwarded by TAO in an ICHN.

A. Assumptions

TAO relies on four main assumptions:

1) Mobile hosts are able to perceive their one-hop neighborhood.
2) Each mobile host is able to temporarily store the messages it receives, and can associate to each of them some pieces of information.
3) All the infostations are continuously running and are connected to each other using a backbone. An infostation is never out of order and can provide services itself, or can act as a proxy to another infostation.
4) The time synchronization between mobile devices does not need to be very accurate, given the low movement speed of the mobile devices, which are carried by pedestrians. We can legitimately consider that TAO can tolerate a clock-shift on the order of half a minute.

B. Overview of TAO

TAO implements the ”store, carry and forward” principle, a multiple-copy message forwarding algorithm, source routing mechanisms and a time-stamping-based heuristic that aims at selecting the best next message forwarder(s) among a set of neighbor nodes. This heuristic relies on lists of last dates of contact of a node with its neighbors. Mobile nodes are expected to use these lists when they are solicited by a neighbor node that wants to forward a message to an infostation or to a given mobile node, and to return to this node their last date of contact with an infostation. These pieces of information are then processed by the mobile node that must forward the message in order to select the best next forwarder(s) to deliver the message to an infostation. Such an algorithm tends to reduce progressively the area where the messages are disseminated until reaching their destination (i.e., a fixed infostation). Indeed, from a logical point of view, a node will disconnect from an infostation when it starts moving away from it, and the distance separating them is increasing progressively.

In order to improve the service delivery and to avoid the bad carrier dilemma (i.e., delivering the message to a mobile node abruptly moving in the wrong way), TAO implements a multiple-copy message forwarding algorithm. The source node is first expected to forward a copy of this message to its best neighbors. Later these carriers, and the source node, will forward a copy of this message only when they encounter a *better carrier* than themselves, while a limited number of copies will be forwarded toward *bad carriers*. A neighbor will be considered as a *good* carrier by a node if the last date of contact of this neighbor with an infostation is more recent than its own date of contact, or if this neighbor is a new neighbor and its last date of contact with an infostation is more recent than those of the other neighbors. This classification aims at estimating the ability of these intermediate nodes to deliver the message to an infostation. Furthermore, sometimes it might be critical to forward copies exclusively to good carriers, especially if the source node is located remotely from an infostation. In this case, the major number of neighbors

 ISBN: 978-1-61208-203-5

might be classified as bad carriers, while the contact with a good carrier is uncertain. Thus in TAO, each mobile node has a stock of a few number of copies dedicated to bad neighbors. This number of copies is limited in order to avoid network overload and resource consumption on mobile devices. Finally, TAO makes it possible to control the propagation of messages in the network using two parameters: a lifetime and a number of hops. When the lifetime is expired or when the number of hops is zero, the message is removed from the local cache and will not be forwarded anymore.

Unlike Prophet [5], PropicMan [6] or HiBOp [7], TAO needs neither to record any large set of history values nor to use complex algorithms to select the next message carriers. In TAO, each node maintains 3 kinds of lists: a neighbor list, a list of last local contacts with infostations and a list of the last remote contacts with the infostations. These 3 lists will be referred to as NL, LLCI and LRCI respectively in the remainder of this section. NL contains the one-hop neighbors of the node and the date of reception of their last beacon. This set is obtained thanks to a background beaconing performed periodically by each node. Furthermore, mobile nodes or infostations are considered disconnected from the neighborhood of a mobile node if no beacons have been received from them during a time gap superior to an already define disconnection time threshold.

The LLCI list contains a limited number of entries (on the order of the number of two-hop neighbors). Each entry includes the IDs of the infostations the node has encountered and the time of the last beacon it received from these infostations. The LRCI list is similar to the LLCI list, but it is related to the neighbor nodes found in the NL list. Each entry of the LRCI represents the ID of the most recent infostation each neighbor had contact with and the value of the last contact date.When a new neighbor joins the one-hop neighborhood of the node it will directly send the ID and the most recent time of contact with an infostation to be registered in the LRCI.

TAO is a reactive and an event-driven protocol. Five events are considered in TAO:

1) The emission of an invocation request by a local client application.
2) The reception of an invocation request sent by a neighbor node.
3) The reception of a service response sent by a neighbor node.
4) The arrival of a new neighbor node.
5) The disappearance of a neighbor node.

C. Forwarding of service invocation requests

When a node receives an invocation request from a local application for a service provided by a remote infostation (event 1), or when it receives such a request from one of its neighbors (event 2), the node will process this request according to the forwarding Algorithm 1. On each message reception, a mobile node will forward a copy of this message to E_{emit} direct neighbors at the most. In this algorithm, the good carriers will always obtain a copy of the message. The infostation contact time of the best carrier of each invocation request will be recorded in the lists of the local node in order to be compared later with those of the new nodes that appear in the neighborhood of the current carrier. A copy of each invocation request will only be forwarded toward a new neighbor if its infostation contact time is more recent than the locally recorded value. As a result, this process prevents from forwarding multiple copies to the same node. On the other hand, bad carriers will receive a copy from the stock of copies dedicated to them only when the number of good carriers is less than E_{emit} in the neighborhood of the carrier node. The stock will be decremented, and when this stock is empty no more copies of the message will be forwarded to the bad carriers.

Algorithm 1 Emission or forwarding of an invocation request.

```
emission or forwarding of a Service Invocation Message :
 if it exists an infostation in the one-hop neighborhood then
   forward request to the infostation
   remove the Service Invocation Message from the cache
 else
   check LRCI list for good carriers
   for number of emissions is less than E_emit
     if good carriers are found then
       forward a copy of the Service Invocation Message
       Update the best contact time with an infostation relative to this
invocation message
     else
       if local stock related to this Service Invocation Message > 0
         forward a copy of the Service Invocation Message
         decrement the stock
       endif
     endif
   endfor
 endif
```

In order to select the best next carriers among its neighbors, the local host checks the content of the LRCI list and chooses the nodes with the most recent infostation contact time. Service invocation messages will be stored, carried and forwarded by intermediate carriers in the same manner.

When a node discovers that it exists in its one-hop neighborhood an infostation, it forwards to this infostation all the service invocation requests it has in its cache and that are still valid, and removes them from its cache of messages. This infostation is expected to either deliver the service itself or forward the request to the appropriate infostations.

Any invocation request copy will stay alive until its lifetime expires or its number of hops is zero. Moreover, each message will store in its header the route it followed to reach the infostation.

D. Forwarding of service responses

The next event is the reception of the reply from the infostation toward the client node. The infostation will send the reply message with a reverse routing, in other words, send the message back on the route it has just traversed. Reverse source routing can be reliable if the mobility of the nodes in the network is relatively slow. In fact, TAO is designed to function in ICHN, where mobile nodes are volatile and end-to-end paths between nodes cannot be maintained all the time.

 ISBN: 978-1-61208-203-5

As a result, a forwarding mechanism must be applied when the reverse source routing fails at some point (i.e., a disconnection in the route is encountered).

Algorithm 2 Management of the forwarding mechanism of the service response.

```
emission of a Service Response Message
if the original-service-requester is connected then
 forward the Service Response Message to the original-service-requester
 remove the Service Response Message from cache
else
 if next id recorded in header is connected (reverse source routing) then
  forward the Service Response Message
 else
  scan NL list for a node of id recorded in the header
  if id found in the neighborhood then
   forward a copy of Service Response Message
   remove the id of the node from the header of the message registered in
the cache
  endif
 endif
endif
```

When such a disconnection occurs, the node carrying the message will perform a multiple-copy message forwarding algorithm dedicated to the service responses (Algorithm 2). A copy of the service response will be forwarded to the best neighbors (setting higher priority to the nodes closer to the destination). Thus, all of the new carriers will try to resume the source routing process. If they cannot resume this process, they forward a copy of the message when they encounter a better carrier than themselves.

E. Management of neighborhood changes

Algorithm 3 Management of the service messages after the contact with a new neighbor.

```
forwarding service messages toward the newly arriving neighbor:
 if event is arrival then
  if the new neighbor is a good carrier or size of stock > 0 then
   foreach valid request in the local cache do
    forward the request to the new neighbor
   endforeach
   if the size of the sock > 0 then
    decrement the stock
   endif
  endif
  foreach valid response in the local cache do
   run algorithm 2
  endforeach
 endif
```

The last two above-mentioned events are triggered when changes occur in the one-hop neighborhood of a node. When a new neighbor joins the one-hop neighborhood of another node, the LLCI and LRCI lists are updated using the pieces of information obtained from the beaconing process. Then, the current carrier is expected to check if it exists in its local cache some messages (requests or responses) that should be delivered. If so, it conditionally forwards copies of these messages to its new neighbor using Algorithm 3. If the new neighbor has a date of contact with an infostation that is more recent than those of the other neighbors of the current carrier, this one will forward to this new neighbor all the service invocation requests it has and that are still valid. The service responses stored locally will be forwarded by the current carrier to this new neighbor only if this one is the destination or has been used as intermediate node to forward the request (i.e., if this one appears in the reverse source path header of the response).

Algorithm 4 Management of the disappearance of a neighbor to the mobile node.

```
disappearance of a neighbor:
 if event is disappearance then
  if the node is an infostation then
   update the LLCI list with the time and the id values of the infostation
included in set NL
  endif
  if the node is a mobile node then
   update the LRCI by eliminating the id and time values of the
respective disconnecting neighbor
  endif
 endif
```

As Algorithm 4 shows, when a node is notified of the disappearance of an infostation from its one-hop neighborhood (i.e., from the neighbor set NL), it updates its LLCI list with the ID and the date of the last beacon of the infostation, that has been removed from the neighbor set. Otherwise, if the disconnected neighbor is a mobile node, all the related information will be removed from the NL and LRCI lists.

III. SIMULATION

In this section, we present the simulation results we obtained for TAO regarding the service delivery performances in comparison with a simple routing protocol RANDOM. The RANDOM routing protocol has the same characteristics as TAO, but instead of relying on the time heuristic implemented in TAO, RANDOM relies on a random mechanism in choosing the next carriers of the service invocation messages. The main objective of these simulations is not to compare the global performance of TAO with other protocols such as [7], [8] and [9], but instead to assess the effectiveness of the time heuristic in delivering messages between a mobile node and a fixed infostation. The simulations have been performed on the OMNeT++ network simulator. In these simulations, we focus on a simple, but realistic, hybrid network composed of only one infostation providing a service and a set of mobile devices carried by pedestrians. A part of these pedestrians act as clients, forming a set of nodes that request services from the infostation.

A. Setup

The simulation environment we consider is a square open area of 1 km^2. The infostation is located in the middle of this

 ISBN: 978-1-61208-203-5

area. All mobile nodes move according to a random waypoint mobility model with a speed of 0.5 m/s. The communication range of the mobile devices and of the infostation is approximately 30 m. After discovering the service they are looking for, the mobile clients invoke this service every 3 minutes. In our experiments, we have not assigned any lifetime or maximum number of hops to any message for both protocols.

We ran the simulations for each routing protocol, varying the number of nodes forming the network. For each setup, we made 10 simulation runs with a different random seed. A *warm up* period of 10 minutes is used in the beginning of the simulations before clients start to generate invocation request messages, to allow LLCI, LRCI and NL lists to be initialized. The simulation is run for another 1 hour before stopping all the invocation requests from all the clients. Finally, the simulation is left for 10 minutes to allow all the messages to be delivered.

We present below the obtained results. The objectives of these experiments was to measure the ability to satisfy the client service delivery with a small number of message copies, where we have fixed the number of copies (size of the stock) to 3 copies for both protocols and the maximum number of emissions E_{emit} to 3.

B. Results

Each of the following graphs contain curves for both TAO and the RANDOM routing protocols, while changing both the number of clients and the number of mobile nodes forming the network. The two metrics we are studying are the satisfaction ratio and the delay. The satisfaction ratio is the percentage of successful service invocation (i.e., the number of invocations for which a client node receives their response from an infostation), while the delay is the total time needed by a successful invocation message to travel from the client node toward an infostation and on the way back toward the same client node.

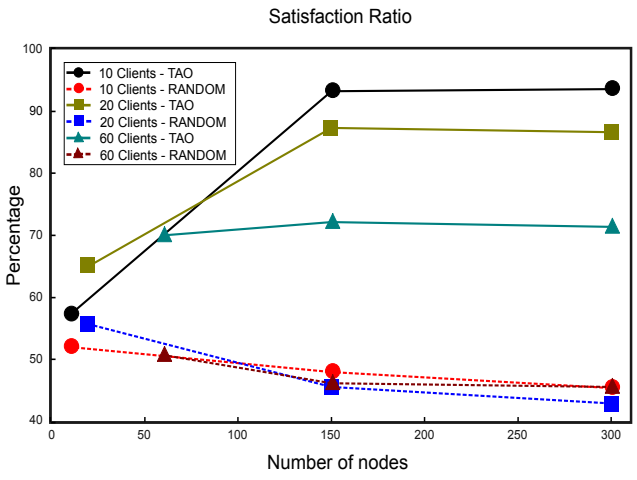


Figure 2. Satisfaction ratios of TAO and RANDOM Routing Protocols

First, we analyze the satisfaction ratio of each of the protocols in order to study the impact of increasing the number of nodes forming the network on the performance of each protocol. Three scenarios for each protocol are presented in Figure 2, where each scenario is characterized by the number of clients found in the network. As we notice, when having few nodes in the network, the satisfaction ratio of both protocols is almost the same. This observation is coherent with what is expected, because, due to the limited number of neighbors, TAO and RANDOM will select most of the time the same carriers. The performance of TAO increases with the number of nodes forming the network due to the selection of good carriers among a large set of neighbors. On the contrary, the performance of RANDOM decreases due to the bad selection of next carriers.

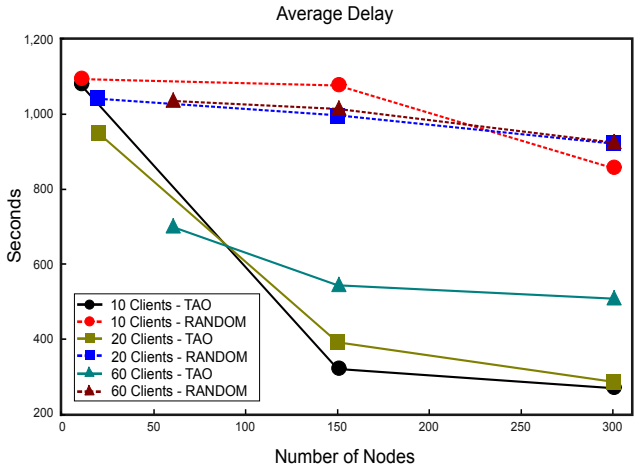


Figure 3. Delay of TAO and RANDOM Routing Protocols

Based on (Figure 3), we notice that when the network is formed of a few number of nodes, the delay values are relatively high. This is totally normal, since few nodes with limited transmission range and random waypoint mobility have to cover a large area. When the number of mobile nodes increases in the network, the average delay of RANDOM remains relatively high due to the random choices of carriers that contribute in transmitting the invocation messages toward the infostations. On the contrary, the average delay of TAO decreases due to the presence of more carriers that can fill the gap between the client and the infostation and the ability of TAO in choosing good carriers to perform this operation. As a result, the mechanism used by TAO to choose among the carriers of service invocation messages results in shorter delays and higher delivery ratio than what RANDOM can offer.

Finally, the application we utilized in these primary simulations was simply a mobile node invoking an infostation and waiting for the response back. This was done to study the effect of the time heuristic implemented in TAO. Since the number of hops and the lifetime message parameters are directly related to the application and the expected delays in the targeted network, we did not specify any limits over these two parameters. So, the number of messages that are disseminated in the network increases continuously. That's why we do not give in this paper details about the network

 ISBN: 978-1-61208-203-5

load and the efficiency of TAO regarding this point precisely. Nevertheless, the measurements we conducted show that the number of messages disseminated by TAO and RANDOM is approximately the same.

IV. Related Work

TAO combines the originality of being designed to support service invocation and to specifically target hybrid intermittently connected MANETs. Indeed, most of the research work concerning intermittently connected MANETs consider networks solely formed of mobile devices and multi-purpose communication. However TAO uses the "store, carry and forward" principle and several techniques that are shared with many protocols in delay-tolerant and opportunistic networking. In the remainder of this section, we present some representative routing protocols.

The Message Ferrying paradigm is a mobility-assisted approach [8]. It introduces non-randomness to the node mobility and exploits it to provide physical connectivity among nodes. Two variations are introduced both with movement constraints. The first is targeting the message ferries, which are special nodes introduced as data carriers toward specific positions to collect data. The second forces the clients to move toward the ferries in order to send and receive messages. Although this approach ensures reliability, it introduces a lot of constraints on the mobility of clients, which is considered disturbing and impractical in real life.

The constraint of intervening with clients' mobility was addressed by other protocols that aimed to benefit from the natural uncontrolled movements of nodes. The most general attempt is the flooding-based routing protocols such as Epidemic Routing [9]. In such types of protocols, messages are flooded in the network and stored by all available neighbor nodes. No precautions are considered to limit the number of messages exchanged and forwarded. Therefore, network congestion is very likely to happen, especially in high density regions.

Some approaches resort to probabilistic solutions [5], [6] and prediction techniques [10] to choose the best relay nodes for each specific message. For routing, these protocols rely on the context and history information in order to compute delivery probabilities and predictabilities. A history-based approach such as the one described in [7] relies on an algorithm to predict the future movements and patterns of the node and route packets according to these predictions. Such algorithms can achieve high efficiency and delivery ratio, but the major drawback is assuming that a node is able to register a large amount of contact history in its buffers, taking into consideration that the majority of the devices forming such networks are small portable devices with limited capabilities and capacities. Moreover, computing histories and predicting movement patterns is a tricky problem, especially in environments composed of numerous mobile devices that move following irregular patterns, such as those held by pedestrians in a city. On the contrary, TAO simply utilizes recent history to predict near future. TAO tends to estimate the distance separating the nodes and the infostations in the network, relying on the disconnection times recorded locally by the mobile nodes. By this approach we aim to reduce the complexity and the resources needed to take correct routing decisions.

V. Conclusion

Intermittently connected hybrid networks is a quiet original perspective that can be exploited to provide nomadic people with a wide access to pervasive services, eliminating the need for any expensive infrastructures, such as those provided by GSM operators.

In this paper, we have proposed a new simple time-aware opportunistic routing protocol, called TAO, devoted to service invocation in ICHN. TAO implements a multiple-copy message forwarding algorithm, source routing mechanisms, and a time-stamping-based heuristic that aims at selecting the best next message forwarder(s) among a set of neighbor nodes.

TAO is a recent work, and its evaluation through real conditions is still in progress. TAO will shortly be included in a service-oriented middleware platform. Furthermore, in the future, we would like to improve the service delivery process by implementing a handover mechanism in the infostations, thus allowing to provide nomadic people with a continuous access to a given service. Then, a service will be delivered all the time to the client by the best infostation, which will be, the most of the time, the nearest infostation.

References

[1] Y. Mahéo and R. Said, "Service Invocation over Content-Based Communication in Disconnected Mobile Ad Hoc Networks," in *24th International Conference on Advanced Information Networking and Applications (AINA'10)*. Perth, Australia: IEEE CS, April 2010, pp. 503–510.

[2] N. Le Sommer, S. Ben Sassi, F. Guidec, and Y. Mahéo, "A Middleware Support for Location-Based Service Discovery and Invocation in Disconnected MANETs," *Studia Informatica Universalis*, vol. 8, no. 3, pp. 71–97, September 2010.

[3] M. Conti and M. Kumar, "Opportunities in Opportunistic Computing," *Computer*, vol. 43, pp. 42–50, 2010.

[4] M. Conti, S. Giordano, M. May, and A. Passarella, "From opportunistic networks to opportunistic computing," *IEEE Communications Magazine*, vol. 48, no. 9, pp. 126–139, September 2010.

[5] A. Lindgren, A. Doria, and O. Schelen, "Probabilistic Routing in Intermittently Connected Networks," in *Proceedings of the The First International Workshop on Service Assurance with Partial and Intermittent Resources (SAPIR 2004)*, Fortaleza, Brazil, August 2004.

[6] H. A. Nguyen, S. Giordano, and A. Puiatti, "Probabilistic Routing Protocol for Intermittently Connected Mobile Ad hoc Network (PROPICMAN)," in *International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2007)*. Helsinky, Finland: IEEE CS, June 2007, pp. 1–6.

[7] C. Boldrini, M. Conti, J. Jacopini, and A. Passarella, "HiBOp: a History Based Routing Protocol for Opportunistic Networks," in *International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2007)*, M. Conti, Ed. Helsinky, Finland: IEEE CS, 2007, pp. 1–12.

[8] W. Zhao, M. Ammar, and E. Zegura, "A Message Ferrying Approach for Data Delivery in Sparse Mobile Ad Hoc Networks," in *Proceedings of ACM Mobihoc 2004*, Tokyo Japan, May 2004.

[9] A. Vahdat and D. Becker, "Epidemic Routing for Partially Connected Ad Hoc Networks," Duke University, Tech. Rep., April 2000.

[10] M. Musolesi and C. Mascolo, "CAR: Context-Aware Adaptive Routing for Delay Tolerant Mobile Networks," *IEEE Transactions on Mobile Computing*, vol. 8, no. 2, pp. 246–260, 2009.

 ISBN: 978-1-61208-203-5

Compulsory Service Extensions to SIP-Initiated Communication Sessions

Philipp Marcus, Thomas Mair, and Florian Dorfmeister
Institut für Informatik
Ludwig-Maximilians-Universität München
Munich, Germany
Email: {philipp.marcus, thomas.mair, florian.dorfmeister}@ifi.lmu.de

***Abstract*—The Session Initiation Protocol is an application layer protocol with increasing impact on signaling in mobile networks and mobile applications. However, it lacks the possibility to enforce the abidance of special communication constraints after session parameters have been negotiated. In this work, an approach of compulsory services is presented, which allows peer user agents in communication sessions set up by the Session Initiation Protocol to mutually prove each other that transmitted data has been processed by a trusted third party. This is realized by enforcing a set of compulsory services implemented by trusted third parties. Technically, in the session initiation phase the signaling messages are therefore modified according to a rule base in special policy servers. During data transmission, an involved user agent first passes its message to the compulsory service, receives a corresponding token as acknowledgment and finally transmits this token along with the message to the peer user agent. There, based on the validity of the token, further actions can be taken. The usability is demonstrated by three examples including location verification of mobile users.**

***Keywords-SIP-Services*; *Trusted Third Party*; *Public-Key-Cryptography*; *Location Verification*; *Call Recording*.**

I. Introduction

The Session Initiation Protocol (SIP) is an application layer protocol developed by the IETF with focus on creating, modifying and terminating communication sessions with one or more participants [1]. It is designed according to design principles similar to HTTP, and therefore is well suited for IP-based networks. It also allows for extending the protocol with new features that not necessarily must be supported by all clients.

Although SIP is primarily used for Internet telephone calls and video conferencing, due to its flexibility it can also be applied to other domains such as instant messaging. In the context of next generation networks and the IP Multimedia Subsystem (IMS) specified by the 3GGP in TS 23.228, SIP is used as a session control layer to provide a standardized interface to services between mobile users and signaling infrastructure.

The architecture of a typical SIP configuration consists of several entities of which the most important in the context of our work comprise *user agents* (UAs) and the signaling infrastructure elements, i.e., *registrars* and *proxies*. One or more user agents register themselves with a SIP username at a *registrar*, which implements a location service mapping usernames to network addresses. For session initiation, a UA sends an `INVITE` message to the peer user agent. Those messages are typically passed through one or more proxies, which are responsible for routing SIP messages and thereby are able to enforce a call policy.

After the session has been established between the involved user agents, the data transmission phase will follow. Here, none of the signaling infrastructure components has further access to the information exchanged. Hence, although desirable in many application scenarios like the enforcement of mutually agreed on centralized call recording or the mutual provision of reliable information about the residence of UAs, the standard SIP provides no means for enforcing certain properties or the processing of transmitted data.

In order to facilitate the enforcement of specific processing constraints during the data transmission phase, we developed an approach that enables user agents to mutually prove that a given piece of transmitted information has been processed by a set of trusted third party services called *compulsory services* in a predefined manner. The set of involved compulsory services is deduced from policy servers and contained in the SIP-negotiation messages. The services themselves are hosted by external providers.

The remainder of the paper is structured as follows: Section II presents the state of the art in form of related work. Section III then describes the principles of compulsory services comprising the theoretical concepts, the structure of compulsory service assignments in the SIP-negotiation phase, the necessary extensions to SIP and discussion of the data transmission protocol. In Section IV, three example scenarios are presented that profit from compulsory services. Finally, Section V concludes the paper.

II. Related Work

SIP services can be classified into being active during the session initiation or focusing on modifying the data transmission phase. During the session initiation, services can run on signaling infrastructure or on user agent equipment depending on the task the service performs and aim at enforcing parameters for the following data transmission phase according to a set of rules [2]. Device independent services, for example services taking actions when systems

 ISBN: 978-1-61208-203-5

are not available or busy, are run on signaling infrastructure, e.g., using SIP-CGI [3]. Services comprising device specific information like location or call waiting alerts are executed on UA equipment. SIP services typically take the following steps:

1) Modification of headers and forwarding of incoming requests based on a set of rules.
2) Receipt of replies to forwarded requests, modifying the replies and returning them to the client.
3) Generating requests to other services by creating and sending appropriate messages.
4) Generation of responses to incoming requests and sending the results to the client.

Approaches to the specification of services that are run on UA equipment have been proposed with SIP Call Processing Language (CPL) [2] or Language For End System Services (LESS) [4]. These approaches might be useful for implementing UAs that dynamically adapt to an incoming set of compulsory services during the session initiation.

A hybrid approach is the concept of SIP Back to Back User Agents (B2BUAs) working as a *man in the middle* in the SIP signaling process and therefore incorporating a UAC and UAS at the same time for the transmission of a given message. A B2BUA keeps track of the state of connections it has established and has full control over all signaling messages [1]. Extensive research has been done in developing domain specific languages for B2BUA programming, enabling applications like seamless device handover [5] [6]. Other approaches like DiaSpec focus on the automatic generation of includable source code from telephony service specifications [7]. In comparison, our approach does not aim at modifying the delivered content but rather extending it with a token certifying that a piece of information in a message has been processed by a trusted third party, namely a compulsory service in a predefined manner.

Services in the data transmission phase have especially been in the focus of research and have been developed in the context of lawful interception (LI) [8]. Focused on the interception of voice transmissions, signaling infrastructure typically modifies `INVITE` and `ACK` messages in the SDP information in order to enforce the redirection of the packet through a special recording gateway [9]. This way, the internet telephony service provider in charge is able to apply logging or packet sniffing tools to the data stream. Contrary to our approach, the employment of the recording services can be concealed from the communicating user agents. Another downside of these approaches is that it is not possible to avoid repudiation of received packages, which in contrast can be implemented by our compulsory services.

III. Compulsory Services and SIP-Sessions

In this section, we present our approach to enforcing the interaction of user agents with compulsory services in SIP initiated communication sessions.

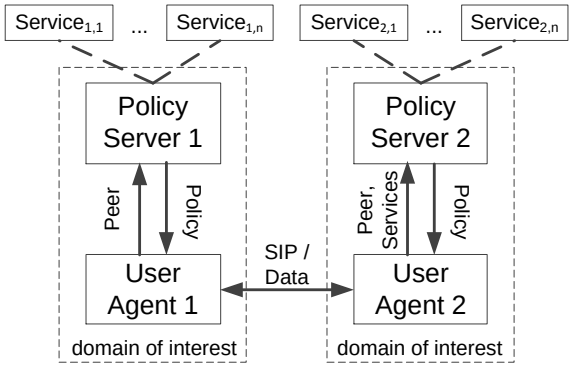


Figure 1. System architecture showing services known by policy servers and two communicating user agents.

A. System Architecture

We propose a system architecture as depicted in Figure 1. A policy server (PS) within an administration domain is associated to $1 \ldots n$ services. Before starting the SIP-negotiation, a user agent queries its PS, which has an integrated set of rules stating which services have been defined as compulsory for the interaction between the two involved user agents. Those services are called *compulsory services* (CS). Conceptually, these services act as trusted third parties, enabling clients to mutually guarantee that data sent to the other user agent has also been sent to and processed by each CS. This is realized by forcing each involved user agent to request *tokens* with a limited time of validity from the assigned compulsory services. These tokens are only issued by the service if the user provides service-specific required data. The tokens are then appended by the UA to data transmissions, which will only be accepted by the receiving user agent if the token is valid. Before initiating a session with a SIP-`INVITE`, a user agent provides its policy server the identities of himself and his peer and receives a set of CSs along with the specification of their communication interface, also including their service access points *SAPs*. The SIP-`INVITE` is then extended by this set and sent to the peer UA, which again contacts his own policy server. The reply sent back to the first UA is extended by the set of CSs that were received from the peer and those received from his own policy server.

As this process runs the risk that two clients might agree on ignoring the tokens and circumvent the enforcement of the compulsory services, our approach is based on the precondition that each involved policy server is within the same domain of interest with one of the user agents, i.e., for a given policy server, at least one user agent insists on abiding its claimed compulsory services.

 ISBN: 978-1-61208-203-5

More precisely, a *compulsory service* i was defined as a tuple: $CS_i = (SAP, X, f, D, Cert, g)$ with the following semantics: SAP is the service access point of the CS represented by an URI. The set X represents the domain of data CS_i is interested in to finally create a token. However, in general this data is not passed to CS_i by an user agent genuinely, but is rather pre-processed by the function $f : X \mapsto D$. For example, D can be defined as the set of all hash values of elements in X. The *Cert* element represents a public-key certificate associating a public key $K^+_{CS_i}$ with the *SAP* and also states that CS_i owns the private key $K^-_{CS_i}$. The set of all encrypted elements of D is denoted as $\mathcal{C}(D) = \{x | \exists y \in D : K^-_{CS_i}(y) = x\}$.

The last element g is statically defined as a constant $g : \mapsto \{\texttt{once}, \texttt{periodicTime}(t), \texttt{periodicPacket}(t), \texttt{always}\}$, formalizing when tokens need to be generated for a data packet. $\texttt{periodicTime}(t)$ states that at least every t seconds a valid token has to be used. Similarly, $\texttt{periodicPacket}(t)$ implies that at least every t packets a new token has to be provided. In Section III-C the necessary extensions to SIP are discussed.

A compulsory service CS_i provides an external interface to user agents with a function $CS_i : D \mapsto \mathcal{C}(D) \cup \{\perp\}$, which generates the desired tokens. The process of token generation within CS_i for a request $CS_i(f(x))$ is based on two steps: first, the argument $f(x)$ is sent to a processing function $p : D \mapsto \{\texttt{true}, \texttt{false}\}$, e.g., for logging $f(x)$ to a database or consulting third parties to verify information. Its result is a boolean value. If this value is `true`, the token is generated by computing the digital signature of $f(x)$ using $K^-_{CS_i}$ and returned to the calling user agent. Otherwise, $\perp$ will be the result, symbolizing invalid tokens:

$$CS_i(f(x)) = \begin{cases} K^-_{CS_i}(f(x)), & \text{if } p(f(x)) = \texttt{true} \\ \perp, & \text{otherwise} \end{cases} \quad (1)$$

Finally, the calling user agent transmits all tokens $1, \ldots, n$ alongside the message x to his peer where the validity of each token is checked:

$$K^-_{CS_i}(f(x)) \text{ is valid} \Leftrightarrow K^+_{CS_i}(K^-_{CS_i}(f(x))) = f(x) \quad (2)$$

The formalism describing which services are defined as compulsory for a communication session is discussed in the next section.

B. Assigning Compulsory Services to SIP-Sessions

In this section, the process of rule selection at a policy server in the domain of interest of one of the two involved user agents is presented. Therefore, sets *Users*, *Roles* and *Services* have been defined. The schema is illustrated in Figure 2.

The set *Users* contains all known user agents and *Roles* all role identifiers. Users can be assigned to roles given by an user-role-relation $UR \subseteq Users \times Roles$. The set of all

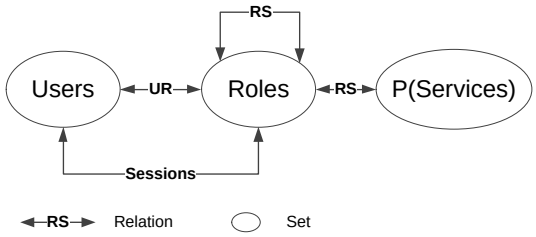


Figure 2. Rule base representing user-role-service-assignments.

currently assigned roles $r \subseteq Roles$ for a user $u \in Users$ is denoted as $r(u)$. Furthermore, the set of all compulsory services has been defined as $Services$ and its power set as $\mathcal{P}(Services)$. The time-dependent role assignment to users at a certain point of time is represented by the set $Sessions \subseteq Users \times Roles$. The actual rule base is defined using the relations $RS \subseteq (Roles \times Roles) \times \mathcal{P}(Services)$ and similarly $\overline{RS}$. RS states, which compulsory services have to be enforced in the communication between two user agents with specific roles, while $\overline{RS}$ contains forbidden ones.

In complex scenarios, each involved user agent, ua_1 and ua_2, has its own policy server PS_1 or PS_2 within its domain of interest. In all cases, ua_1 generates a SIP-`INVITE` according to the rules imposed by PS_1. For this purpose PS_1 provides a set of compulsory services S_{ua_1} that will be integrated into the handshake message. Additionally, ua_2 has the policy server PS_2 within its domain of interest. The latter will will be provided with S_{ua_1} as well as ua_1 and will return a set S_{ua_2} of desired compulsory services and a boolean value indicating if all services in S_{ua_1} are acceptable. In accordance to the extensions made to the original SIP-`INVITE` by ua_1, ua_2 will extend the reply with his set of compulsory services S_{ua_2} induced by the rule base of his own policy server PS_2, too.

A policy server PS_i within the domain of interest of a user agent ua_i computes the set of compulsory services S_{ua_i} according to its rule base by evaluating:

$$\begin{aligned} S_{ua_i} = \{ & cs \in Services | \exists S \in \mathcal{P}(Services) \\ & . \exists r_1, r_2 \in Roles : r_1 \in r(ua_1) \\ & \wedge r_2 \in r(ua_2) \wedge (r_1, r_2, S) \in RS \wedge cs \in S \} \end{aligned} \quad (3)$$

The elements in the set of services S_{ua_k} received in a handshake message from a peer are tested for policy conflicts using $\overline{RS}$. In case of conflicting policies, the session initiation is canceled.

Eventually, each involved user agent knows the union set of all enforced compulsory services $S_{ua_1,ua_2} = S_{ua_1} \cup S_{ua_2}$, which is fixed for the duration of the session.

The next section describes how SIP-messages are extended in order to allow for the transmission of S_{ua_1,ua_2} to the user agents.

 ISBN: 978-1-61208-203-5

C. Integration into SIP

In order to agree on a set of compulsory services between two user agents, the SIP messages at the initiation of a session need to carry information about these services. This guarantees that the user agents are able to reject the session establishment in case one of the user agents does not accept the compulsory services required by the other party.

The inclusion of additional information into a SIP-message may be achieved by different means. Possibilities are to define an additional header field, to send the information within the body of the message or to use multipart bodies when the SIP message already contains a body part (e.g., an SDP body) [10]. Making use of the message body to include the additional information is reasonable if the additional information is extensive and contains structured information. If a user agent encounters an unknown header field, the header will simply be ignored. This solution has the advantage that a user agent that is not able to process new header fields may be detected to be incompatible. If a user agent is not able to decode multipart messages it will issue a `415 Unsupported Media Type` response to signal its inability to process the message. In this work, to add the information, adding the required compulsory services to the body of the SIP-messages has been chosen. This requires only little modification of established user agents and allows for great flexibility when implementing a suitable container format for the description of the compulsory services. The definition of the container format is omitted here, but it will include all the information contained in the definition of a compulsory service as stated in Section III-A.

Each user agent follows the same steps when a session is established as depicted in Figure 3. The calling user agent ua_1 requests the set of compulsory services S_{ua_1} from its policy server PS_1 by providing it with the identity of the desired peer. It then includes the received services S_{ua_1} in the SIP-`INVITE` message, which will be forwarded to the called user agent ua_2. The called user agent first examines if the certificates of the compulsory services are valid and then checks if it is able to provide the necessary information for all services. If one of these checks fails, the `INVITE` request will be rejected by generating a `406 Not Acceptable` response. If the services pass this validation, user agent ua_2 – analogous to ua_1 – queries PS_2 for the required compulsory services for a session with ua_1 and whether the services contained in S_{ua_1} are acceptable. If so, it includes S_{ua_2} into the `200 OK` response or sends a `406 Not Acceptable` otherwise. User agent ua_1 will perform a validation of the services required by ua_2, too. If the validation fails, the session will be terminated by an immediate `BYE` request. Otherwise the session will continue with the data transmission phase as described in Section III-D.

SIP is also able to modify the parameters of an active session, which is achieved by the reinvite mechanism of the protocol. This will be used when a compulsory service is no longer reachable or other problems with the services occur. The re-invitation will renegotiate all the compulsory services for each user agent and terminate the session in the same manner as an initial `INVITE` in case the services do not pass the verification steps at each user agent. This gives the services' administrators the ability to change or remove a service without interrupting the data transmission phase and without termination of the session. If the re-invitation was not successful, i.e., one of the compulsory services is not reachable or the user agent is not able to provide all the necessary information for one of the services, the session is terminated by issuing a `BYE` request.

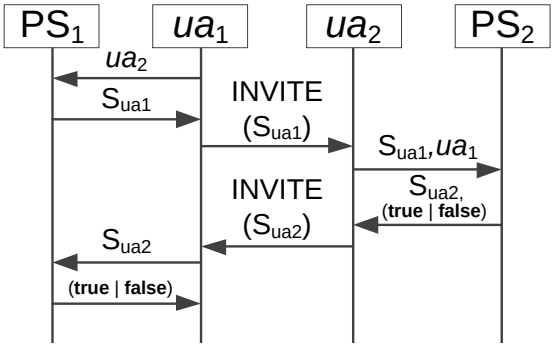


Figure 3. Communication during the the SIP session establishment.

D. Data Transmission Protocol

With the extended session initiation being completed, user agents ua_1 and ua_2 can start their communication as shown in Figure 4. Before actually sending a message x to its peer, however, each user agent has to contact all services contained in S_{ua_1,ua_2} first in order to have its message's content processed and signed by the respective compulsory services. Consequently, the unavailability of any CS leads to the abortion of a session. Since different services might require different representations of x, the requesting user agent ua_1 will locally perform a service-specific pre-processing function f_{CS_i} and provide CS_i with $f_{CS_i}(x)$. Depending on whether a service's processing function p_i works as an idempotent call or not, making use of a reliable communication protocol (e.g., TCP) between user agents and CS_i is required. This is in order to prevent inconsistencies caused by the repeated processing of a message $f_{CS_i}(x)$ due to a lossy channel. CS_i replies to requests with a token $K^{-}_{CS_i}(f(x))$ that ua_1 will then forward to ua_2 alongside the other services' tokens and the original message x.

We propose two alternative solutions for this kind of extended data transmission: The first one is to define a minimal container format bundling the set of required to-

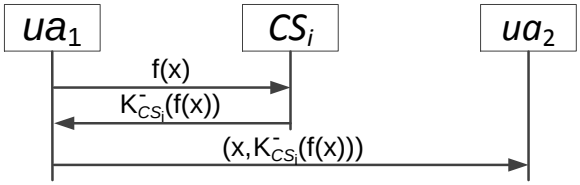


Figure 4. Compulsory service integration in the data transmission phase.

kens $K^-_{CS_i}(f(x))$ and the original message x into a single packet. This tight coupling of a message's content and its associated tokens within one packet is necessary in order to avoid any inconsistencies that might otherwise arise from the loss of packages containing messages or tokens when using unreliable communication channels. Another possible solution to this problem is to embed this kind of compulsory service related information into existing protocols. In case of RTP this could for example be accomplished by assigning a new dynamic payload type `PT` in the RTP header [11] [12]. This allows to avoid defining a wholly new message format by making a minor extension to a standardized protocol a multitude of clients is already speaking. Hence, making use of the latter approach is recommended whenever it seems feasible. Either way, just as desired, only those clients that can either handle such a specific container format or can cope with an existing protocol's extension will be able to correctly process and validate incoming packets as well as give adequate responses.

Upon receiving a message from ua_1 user agent ua_2 will decompose the received packet and then try to validate the contained tokens in order to locally prove that service CS_i has processed and approved the content of x. If any of the tokens required by S_{ua_1,ua_2} are found to be absent or illegal, the discovering user agent will terminate the current session or at least reject the fraudulent message(s).

Dependent on the service-specific value of g_{CS_i}, both the sending and the receiving user agent know when to request and accordingly expect a fresh token from service CS_i. The constant value of g thereby indirectly synchronizes the two user agents' expectations of when an updated token signed by CS is required. Each time one of the agreed on compulsory services demands the usage of a new token, the sending user agent has to repeat the procedure described at the beginning of this section in order to acquire a fresh token. Otherwise the peer client will reject incoming packages until a valid token arrives or even quit the session.

As already stated before, it has been defined as a pre-condition that each user agent ua_i at least abides by the rules imposed by the policy server PS_i that it is assigned to. This is required since the policy servers have no means for supervising the correct usage of compulsory services once the peer-to-peer communication session between the two user agents has been established. Hence, full abidance of the whole set of negotiated services has been achieved by having ua_1 and ua_2 controlling the rules introduced by PS_1 and PS_2 respectively.

IV. Example Scenarios

In this section, three example scenarios and use cases are presented that can be realized using compulsory services.

Location Verification: Especially for mobile user agents the current location of the peer user agent could be of high interest but may be tampered with. Location Verification is the process of proving positions claimed by users [13]. In the context of SIP-sessions this can be realized by implementing a location prover as a compulsory service. In general, a location prover is able to verify the claimed location of a given user agent or entity, e.g., by employing trusted infrastructure like ultrasound sensors in rooms. Assume two communicating user agents that intend to exchange information x along with a verified position l, describing the residence of the sending UA. This results in a message structure $m = (x, l)$. Furthermore, let CS be a location verification compulsory service that has been assigned to the session. Given the message $m = (x, l)$, the pre-processing function f has to be defined as: $f(m) = l$, i.e., only the location l will be transmitted to CS. There, the processing function p will return `true` if the claimed location could be verified and `false` otherwise.

The correctness of l is checked and confirmed with a token $K^-_{CS}(l)$ or with $\perp$ if l does not correspond to the residence of ua_i. According to the delay that is caused by CS when contacting the location proving infrastructure, CS is defined with an appropriate periodicity g. Finally, after checking the correctness of the received token, the peer user agent applies the same processing steps when sending messages to ua_i.

Non-Repudiation Receive: In order to prevent two communicating UAs ua_1 and ua_2 from repudiating the receipt of a message m, a non-repudiation compulsory service CS is assigned to the session. The service has a pre-processing function $f(m, K^-_{ua_i}) = K^-_{ua_i}(hash(x))$, which is used to digitally sign the hash value of a previously received message m. The processing function p is then implemented as the logging of $K^-_{CS}(hash(m))$ to an attached database. This way, CS is able to generate a log of all messages that have been received by the user agents ua_1 and ua_2. The generated token $K^-_{CS}(K^-_{ua_i}(hash(m)))$ is finally sent to the peer user agent by piggy-backing it on a message x in order to acknowledge the receipt of m, thereby preventing the UA from repudiating it later.

Interception and Recording: The assignment of compulsory services to SIP-Sessions can also be used for complete logging of communication sessions on behalf and on the agreement of the user agents ua_1 and ua_2. In many countries, mutual agreement on the recording of calls is claimed by law. The interception or recording of all pieces

 ISBN: 978-1-61208-203-5

of transmitted information is generally more suitable for non real time applications, e.g., instant messaging based on SIP. The periodic requests for attached compulsory services $S_{ua_1,ua_2} = \{CS_1, \ldots, CS_n\}$ introduce a delay d given by the maximum round trip time (RTT) between user agents and compulsory services:

$$d = max\left(\{t | \exists s \in S_{ua_1,ua_2} : t = RTT(s, ua_1) \vee t = RTT(s, ua_2)\}\right) \quad (4)$$

An appropriate compulsory service CS would define a pre-processing function f corresponding to the identity function, i.e., $f(x) = x$. The processing function p here returns `true` if the received data x could be recorded. This is finally confirmed by a ticket $K^{-}_{CS}(x)$, which is sent along with x to the receiving user agent. There, if the token is found to be invalid, stating that x has not been correctly recorded, the processing of x can be stopped.

V. Conclusion and Future Work

We presented an approach that enables SIP user agents to mutually force each other after session initiation to prove that sent data has been additionally processed by a trusted third party in advance, which has been denoted as a compulsory service. During session initiation using an extension of SIP the compulsory services are negotiated, each pre-defined by a pre-processing function and a processing function. A user agent aiming for sending information to its peer first transmits the pre-processed information to all compulsory services and receives individual tokens certifying this interaction. These tokens are finally passed to the peer user agent alongside the original information. During the initiation of a session, the involved user agents extend signaling messages according to a set of compulsory services that are in each case deduced from an attached policy server. This policy server is defined within the same domain of interest of the user agent. We outlined how SIP signaling messages can be extended with a set of compulsory services and discussed aspects of communication between user agents and compulsory services as well as the interaction between two user agents. Finally three applications of compulsory services in the fields of location verification, avoidance of repudiation of message receipt and agreement upon centralized call recording have been presented.

Future work has to concentrate on developing a description language for compulsory services, on optimizing the piggybacking mechanisms for user tokens and evaluating aspects of efficiency.

References

[1] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: Session Initiation Protocol," RFC 3261 (Proposed Standard), Internet Engineering Task Force, Jun. 2002, http://www.ietf.org/rfc/rfc3261.txt (retrieved: April, 2012).

[2] J. Lennox and H. Schulzrinne, "Call Processing Language Framework and Requirements," RFC 2824 (Informational), Internet Engineering Task Force, May 2000, http://www.ietf.org/rfc/rfc2824.txt (retrieved: April, 2012).

[3] J. Lennox, H. Schulzrinne, and J. Rosenberg, "Common Gateway Interface for SIP," RFC 3050 (Informational), Internet Engineering Task Force, Jan. 2001, http://www.ietf.org/rfc/rfc3050.txt (retrieved: April, 2012).

[4] X. Wu and H. Schulzrinne, "Programmable End System Services Using SIP," in *IEEE International Conference on Communications, 2003*, ser. ICC'03, vol. 2. IEEE, 2003, pp. 789–793.

[5] P. Zave, E. Cheung, G. Bond, and T. Smith, "Abstractions for Programming SIP Back-to-Back User Agents," in *Proceedings of the 3rd International Conference on Principles, Systems and Applications of IP Telecommunications*, ser. IPTComm '09, ACM. ACM, 2009, pp. 11:1–11:12.

[6] V. Hilt and M. Hofmann, "Approaches to Implementing Services in SIP Networks," *Bell Labs Technical Journal*, vol. 9, no. 3, pp. 39–44, 2004.

[7] W. Jouve, N. Palix, C. Consel, and P. Kadionik, "A SIP-Based Programming Framework for Advanced Telephony Applications," in *Principles, Systems and Applications of IP Telecommunications. Services and Security for Next Generation Networks*, ser. Lecture Notes in Computer Science. Springer, 2008, vol. 5310, pp. 1–20.

[8] W. Chen, C. Gan, and Y. Lin, "NTP VoIP Platform: A SIP VoIP Platform and Its Services," in *Proceedings of the 5th WSEAS International Conference on Applied Computer Science*, ser. ACOS'06, World Scientific and Engineering Academy and Society (WSEAS). World Scientific and Engineering Academy and Society (WSEAS), 2006, pp. 756–761.

[9] M. Handley, V. Jacobson, and C. Perkins, "SDP: Session Description Protocol," RFC 4566 (Proposed Standard), Internet Engineering Task Force, Jul. 2006, accessed 12-April-2012. [Online]. Available: http://www.ietf.org/rfc/rfc4566.txt(retrieved:April,2012)

[10] G. Camarillo, "Message Body Handling in the Session Initiation Protocol (SIP)," RFC 5621 (Proposed Standard), Internet Engineering Task Force, Sep. 2009, http://www.ietf.org/rfc/rfc5621.txt (retrieved: April, 2012).

[11] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," RFC 3550 (Standard), Internet Engineering Task Force, Jul. 2003, http://www.ietf.org/rfc/rfc3550.txt (retrieved: April, 2012).

[12] H. Schulzrinne and S. Casner, "RTP Profile for Audio and Video Conferences with Minimal Control," RFC 3551 (Standard), Internet Engineering Task Force, Jul. 2003, http://www.ietf.org/rfc/rfc3551.txt (retrieved: April, 2012).

[13] N. Sastry, U. Shankar, and D. Wagner, "Secure verification of location claims," in *Proceedings of the 2nd ACM Workshop on Wireless Security*, ser. WiSe '03. ACM, 2003, pp. 1–10.

 ISBN: 978-1-61208-203-5

Constrained Priority Countdown Freezing - A Collision Memory Avoidance Algorithm

Ivan Kedzo
University Department of Professional Studies
University of Split
Split, Croatia
e-mail: ikedzo@oss.unist.hr

Julije Ozegovic, Vesna Pekic
Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture
University of Split
Split, Croatia
e-mail: julije.ozegovic@fesb.hr
e-mail: vesna.pekic@fesb.hr

***Abstract*— Collision memory in IEEE 802.11 Distributed Coordination Function (DCF) has been detected. The collision memory can increase the physical collision rate and this effect is inherent to any DCF type of countdown. In this paper, we introduce a collision memory avoidance algorithm, called Constrained Priority Countdown Freezing (CPCF). The CPCF can completely or partially remove collision memory depending on how many priority freezing steps are allowed. Since DCF's well known countdown decreases the contention overhead, but increases collision memory effect, the solution is to find the compromise between the two, in order to achieve good performance in both low and high load network conditions. The CPCF achieves this by limiting the countdown process, and thus reducing the collision memory, while still producing significant countdown effect.**

Keywords – collision memory; constrained freezing; backoff freezing; DCF countdown; wireless MAC.

I. INTRODUCTION

The basic channel access method for wireless networks with distributed access is IEEE 802.11 Distributed Coordination Function (DCF) [1]. DCF has one distinguished feature called backoff freezing mechanism, which allows count down of priority through multiple Contention Resolution Periods (CRP), where priority is chosen from certain Contention Window (*CW*). Freezing the countdown process after loosing the medium contention is the key component of DCF, which insures shorter contention overhead. This reduction of contention overhead is called countdown effect [2][6].

However, in our previous work [3], we have seen that countdown freezing mechanism has impact on collision rate. In [3], it can be observed that DCF-like countdown protocol called *Binary Priority Countdown – DCF Countdown* (BPC-DC), which is essentially binary countdown version of DCF protocol, exhibits increased collision rate when compared to *Binary Priority Countdown – Decrement After LCI only* (BPC-DAL) protocol, which chooses new priorities before each CRP. In [3], it is concluded that the reason for increased collision rate is collision memory.

Collision memory occurs due to freezing of the priority countdown, which, as a result, preserves priority collisions. If the freezing of the countdown is not constrained, like in DCF, priority collision always leads to physical collision. This can increase physical collision rate and this effect is called collision memory effect.

In this paper, we propose collision memory avoidance algorithm called Constrained Priority Countdown Freezing (CPCF). CPCF protocol puts constrains on freezing mechanism of DCF in order to achieve short contention overhead and low collision rate.

The rest of the paper is organized as follows. In Section II, the related work is presented. Section III explains in details collision memory, while in Section IV we introduce a new collision memory avoidance algorithm called CPCF. Section V verifies proposed protocol using the ns-2.33 simulator and comments on simulations results. In Section VI, we consider future work and conclude.

II. RELATED WORK

Medium Access Control (MAC) protocols which use Priority Number (*PN*) to resolve the medium contention are called priority contention protocols. These protocols schedule competing Stations (STA) regarding their priorities, allowing higher priority competitors to access the medium earlier. Priority number *PN* is chosen from the set of allowed values called Priority Space (*PS*). If a priority contention protocol employs priority countdown, then lower *PN* should indicate higher priority (e.g., IEEE 802.11 DCF protocol).

DCF uses priority number *PN* as the number of consecutive time slots in which STAs have to wait before starting transmission to the medium. DCF requires a STA to calculate Backoff Counter (*BC*), which is essentially a priority number *PN*, after each transmission. *BC* is chosen randomly from the priority space *PS* limited with the *CW*. After the channel is sensed to be idle for a Distributed Inter Frame Space (DIFS) interval, a STA decrements *BC* when the medium is idle in the current time slot, and *BC* is frozen when another STA is transmitting. When *BC* is decremented to zero, STA accesses the medium.

 ISBN: 978-1-61208-203-5

The 802.11 DCF function has been excessively studied. This included different analysis and enhancements in order to explain or fix DCF's drawbacks. A new protocol, called Enhanced DCF (EDCF), was introduced, supporting Quality of Service (QoS), and it became a new standard [1]. Also, various enhancements were proposed to increase throughput or influence fairness or delay [7][8].

The throughput increase is done mainly through *CW* [7] adaptations in order to reduce collisions or contention overhead, while different Inter Frame Space (IFS) values are used to achieve fairness and low delay for different types of traffic [1].

III. COLLISION MEMORY

DCF countdown allows unconstrained priority freezing after losing the medium contention. The priority chosen by the STA that has lost the medium contention is counted down through multiple CRPs, until it reaches the highest priority and either wins the medium access or enters the collision. When two or more STAs choose the same priority in CRP, priority collision occurs. If the priority collision occurs between the highest priorities chosen, the physical collision occurs in the observed CRP. Obviously, in a single contention there can be multiple priority collisions without any physical collision.

Collision memory can be defined as the ability to preserve priority collisions from previous CRPs, which have occurred during the countdown freezing process. Collision memory can increase the physical collision rate and this effect is called collision memory effect. The collision memory effect occurs due to countdown freezing mechanism. After losing the medium contention, STAs decrement their priority numbers *PN* with the winning priority from the current CRP and freeze decremented *PN* values to be used in the next CRP. All STAs that have experienced a priority collision, will remain in priority collision in the next CRP, and can potentially cause physical collisions.

Collision memory can be preserved through one or more CRPs, depending on the protocol's "memory size". For instance, DCF has infinite collision memory since it can theoretically freeze countdown indefinitely. Remembering priority collisions from previous CRPs is a major drawback when combined with countdown mechanism, which increases STA's priority after each CRP. The DCF's unconstrained priority countdown with freezing insures that all priority collisions eventually become collisions with the highest priority and therefore cause physical collisions.

This can be avoided if we constrain the DCF's backoff procedure. In [3], it is shown that protocol without countdown freezing mechanism, can achieve lower collision rate. However, such protocol also shows increased contention overhead due to lack of the DCF's countdown effect. Therefore, deep investigation of DCF's countdown mechanism with freezing is important, if both, low collision rate and short contention overhead, are desired.

IV. COLLISION MEMORY AVOIDANCE ALGORITHM

The reason causing collision memory effect, which can decrease overall throughput, is DCF's unconstrained priority countdown freezing mechanism. Let's consider the following formula:

$$P_{CM}(m) = \sum_{i=1}^{m} p_i \qquad (1)$$

$$m = 1,\ldots,\infty$$

where $P_{CM}(m)$ is the probability that a STA has experienced the priority collision due to collision memory, after it has frozen its priority m times. Since collisions are remembered due to freezing of priority, $P_{CM}(m)$ is equal to the sum of m probabilities denoted with p_i, where each p_i represents the probability that a STA has experienced the priority collision in i-th step of priority freezing. From this formula it is clear that the bigger m we have, the higher probability $P_{CM}(m)$ becomes. Obviously, the $P_{CM}(m)$ probability has direct influence on physical collisions. STAs that are in countdown can win the medium contention after several consecutive CRPs, and have higher chance of experiencing the physical collision due to high $P_{CM}(m)$ probability.

Therefore, in order to reduce collision memory effect, the priority freezing can be constrained in a way that we can control how many times can priority be frozen and decremented before choosing a new priority. We call this collision memory avoidance algorithm, a Constrained Priority Countdown Freezing (CPCF).

CPCF STA has a counter called freezing counter and a countdown freezing limit k. Parameter k defines the maximum number of times we can freeze and countdown priority, before choosing a new priority value. The parameter k is used to limit the actual number of countdowns m from (1):

$$m = 1,\ldots,k \qquad (2)$$

In (2), m is constrained, and the maximum number of CRPs in which a STA can countdown its priority is equal to k. Besides partially constraining countdown freezing, CPCF can also completely remove it by not allowing STAs to freeze their priorities. This is done by setting the freezing limit k to zero (k=0), which forces STAs to choose new random priorities in each CRP. In this case the $P_{CM}(m)$ is equal to zero, since STAs have no memory of priority collisions that occurred in the past.

The algorithm works as follows. In the beginning, all STAs choose their priorities randomly from certain *CW*, and reset their freezing counters to freezing limit k. CPCF STAs that have lost the medium contention will decrement freezing counter by one, and decrement their priority with the winning priority from the current CRP, just like in

 ISBN: 978-1-61208-203-5

DCF. When freezing counter becomes zero, a STA must choose a new priority number and reset freezing counter to freezing limit *k*. This way, STAs that have lost the medium contention can countdown their priority at most through *k* consecutive CRPs.

A STA that has won the medium also chooses the new priority number, and resets the freezing counter to *k*. After collision, a STA doubles its *CW*. Initial value is set to CW_{min}, and can be increased until it reaches the CW_{max}. This mechanism is identical to the DCF's Binary Exponential Backoff (BEB) mechanism [4]. Obviously, DCF countdown mechanism is CPCF mechanism with $k=\infty$.

V. SIMULATIONS

The performance of the proposed CPCF protocol is verified using the network simulator ns2, version 2.33. The simulator is upgraded with the CPCF module based on ns2 mac-802_11Ext module [5]. The main performance measure is the network throughput achieved, whereas collision probability graphs are presented for reference. For comparison with CPCF, a basic 802.11 DCF MAC protocol was used. Simulations include verification of CPCF's collision memory avoidance algorithm using different values of *k*.

In the network scenario used, a simple wireless ad-hoc network where all *n* STAs can hear each other is simulated. STAs positions are fixed and chosen at the beginning of the simulation, with STAs randomly choosing coordinates from the predefined area. Each STA with address *a* has one ftp flow (bulk packet transfer) directed towards the STA with address (*a*+1)/modulo *n*. Flows are started gradually, from the beginning of the simulation, every 0.1s. Ftp flow is carried over tcp (tcp receiver window is 20 packets wide). Two sets of scenarios are simulated, with tcp packet sizes set to 250 bytes in one, and 2000 bytes in the other set. No MAC segmentation is used and capture effect is turned off. The number of active STAs is increased gradually from 2 to 20, simulating the most frequent numbers of STAs in actual ad-hoc networks. Simulations are repeated with different *k* and CW_{min} parameters. In addition to CPCF parameter *k*, other Physical Layer (PHY) and MAC parameters used are inherited from ns2 802_11Ext class. Table 1 shows fixed parameters used in all simulations.

PHY bandwidth of 6 Mbps is chosen to emphasize the influence of packet lengths used in simulations (250 and 2000 bytes). Low PHY bandwidth produces greater ratio between packet transmission time and overhead than high PHY transfer rates. This way, overall throughput is less affected by overhead, and more by transferring of large collided frames. Therefore, when packets are large (2000 bytes) and low PHY bandwidth is used, collision rate has more influence on throughput than would have for faster PHY. Short packets (250 bytes) represent the real-time traffic and have smaller aforementioned ratio, so keeping the contention resolution period short becomes more important, even with low PHY bandwidth.

In simulations, both DCF and CPCF use three different CW_{min} values (15, 31 and 63). For each CW_{min} value, four different freezing limits *k* are used (0, 1, 2 and 6).

Figures 1, 2 and 3 show the throughput results when both protocols use CW_{min}=15, CW_{min}=31 and CW_{min}=63, respectively.

TABLE I. FIXED CPCF PARAMETERS

Parameter	Value
SIFS	16 μs
Slot Time	9 μs
ShortRetryLimit	7
RTSThreshold	3000 bytes
PHY bandwidth	6 Mbps

DCF countdown achieves good results when the number of STAs is low (up to 4) due to countdown effect. The benefit of DCF's countdown effect is especially visible in Figure 2a and 3a where CW_{min} is 31 and 63, respectively. When CW_{min} insures low collision rate, reducing the contention overhead becomes extremely important, especially when packets are short (collision loses are less expensive in terms of throughput), and thus DCF achieves better results than CPCF. When the number of STAs becomes large, DCF shows poor performance. This can be explained with collision memory, which increases overall collision rate when the number of STAs increases.

CPCF protocol shows good performance, in both low and high load conditions, depending on the freezing limit *k* used. For large *k* value (*k*=6), CPCF has the ability to countdown longer and can significantly reduce the contention overhead. This is very important when the number of collisions is low (STA count is low). However, long countdown can increase collisions due to collision memory and this is the reason why *k*=0, *k*=1 show better performance when the number of STAs gets large.

The really interesting effect occurs when the number of STAs becomes very large (above 15). Since the smallest parameter *k*=0 insures no collision memory effect, it was expected that it would produce the lowest collision rate. However, this was not true. When the number of STAs gets large, *k*=0 and *k*=1 graphs show small but definite difference in collision rate in favor of *k*=1 (Figures 4, 5 and 6). There is obviously another mechanism affecting collisions besides collision memory.

One possible explanation can be found in disturbed distribution of *PN* choices because of constrained countdown freezing. This introduces complex relationship between contention overhead and collision probability. This is visible in the Figure 3a and 3b when the number of STAs is 20. CPCF *k*=0 exhibits better throughput results for short packets than for long packets, when compared with *k*=1. Surprisingly, *k*=0 achieves good throughput results due to shorter contention overhead and not due to lower collision rate as expected. This is confirmed in Figure 6, where it is visible that the *k*=1 shows lower collision rate than *k*=0. Therefore, constrained countdown

 ISBN: 978-1-61208-203-5

freezing mechanism should be further investigated in the future.

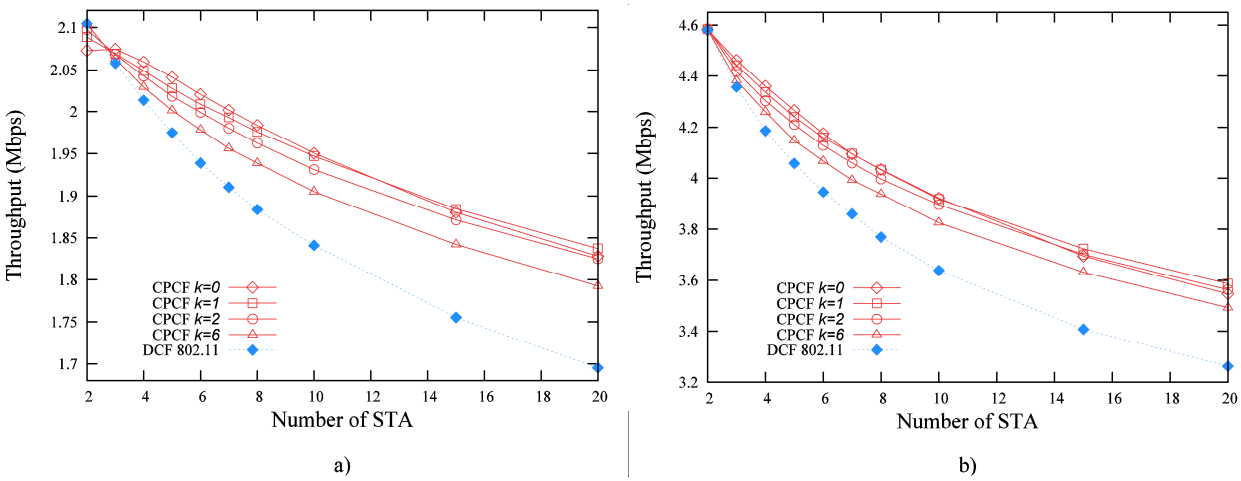


Figure 1. Throughput CW_{min} =15: (a) tcp packet size 250 bytes (b) tcp packet size 2000 bytes

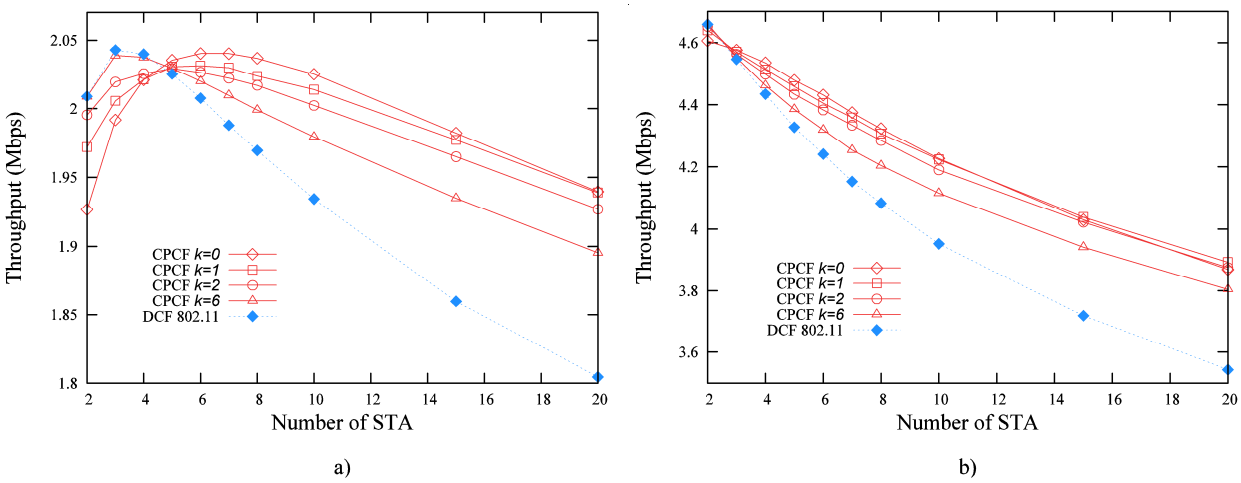


Figure 2. Throughput CW_{min} =31: (a) tcp packet size 250 bytes (b) tcp packet size 2000 bytes

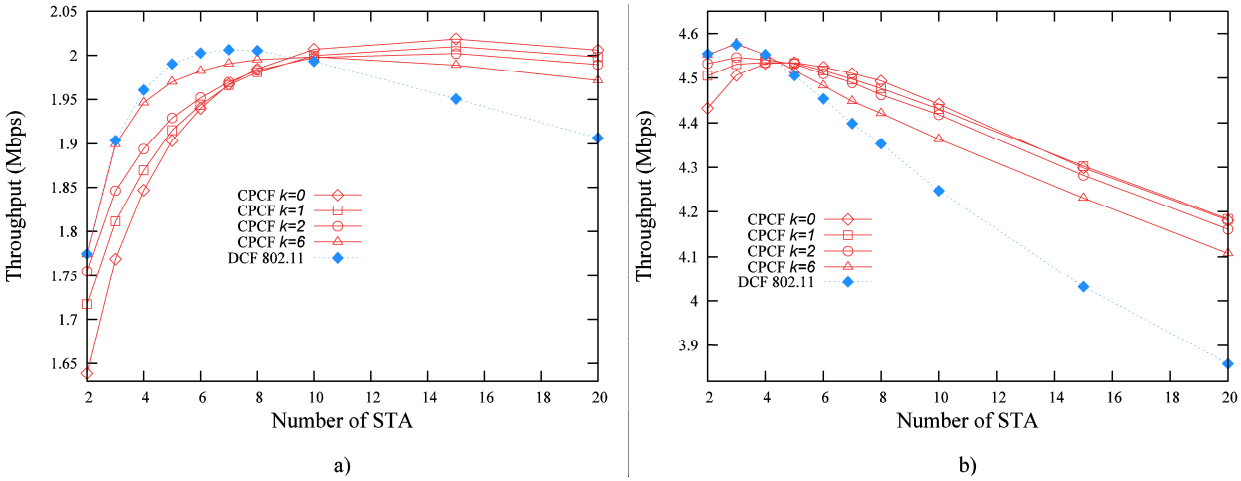


Figure 3. Throughput CW_{min} =63: (a) tcp packet size 250 bytes (b) tcp packet size 2000 bytes

 ISBN: 978-1-61208-203-5

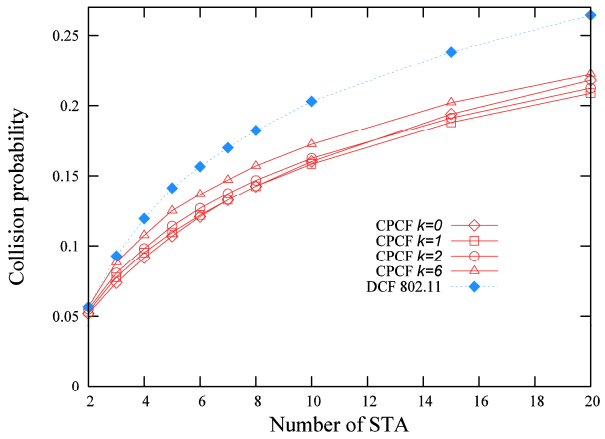


Figure 4. Collision probability CW_{min} =15

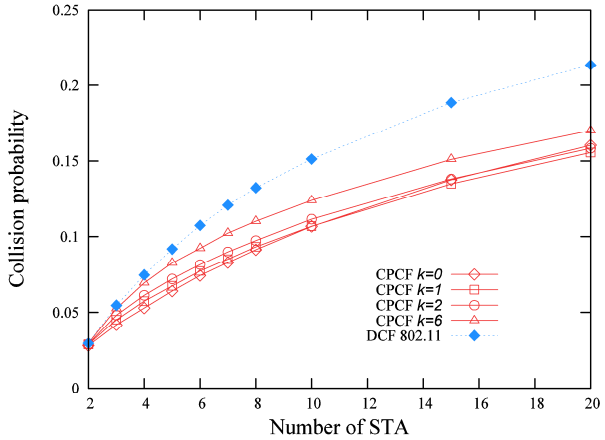


Figure 5. Collision probability CW_{min} =31

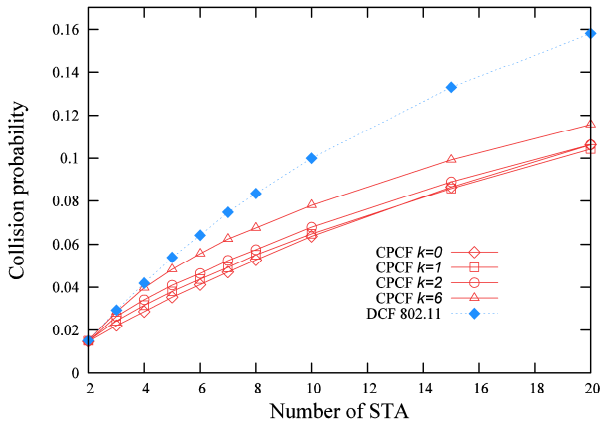


Figure 6. Collision probability CW_{min} =63

VI. CONCLUSION AND FUTURE WORK

In this paper, we have elaborated collision memory and have proposed a new protocol that can reduce collision memory effect. The backoff freezing mechanism of DCF protocol causes collision memory effect by preserving once developed priority collisions, which can increase the physical collision rate. In order to solve the problem, a new collision memory avoidance protocol called Constrained Priority Countdown Freezing (CPCF), is introduced. The CPCF constrains the priority freezing with freezing limit *k*. Freezing limit defines the maximum number of contention resolution periods in which a STA is allowed to decrement and freeze its priority. Simulations have shown that by constraining the countdown freezing mechanism, better throughput results are achieved compared to DCF protocol. However, the simulations have also shown that countdown effect and collision memory effect are not the only effects occurring in CPCF type of countdown. Surprisingly, the most constrained version of CPCF protocol (when k=0), which has no collision memory, can have higher collision rate and lower contention overhead compared to less constrained versions (k=1, k=2). An investigation of this effect should be subject of the future work.

REFERENCES

[1] IEEE Std. 802.11, "IEEE Std. 802.11-2007, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications," 2007.

[2] S. Choi, K. Park, and C. Kim, "On the performance characteristics of WLANs: revisited," ACM SIGMETRICS Performance Evaluation Review, ACM, 2005, pp. 97-108.

[3] I. Kedzo, J. Ozegovic, and A. Kristic, "BPC – a binary priority countdown protocol," *Ad Hoc Networks*, submitted for publication, unpublished.

[4] IEEE 802.11 WG, "Wireless LAN Medium Access Control (MAC) and Physical-Layer (PHY) Specifications," *IEEE std*, 1999.

[5] Q. Chen, F. Schmidt-Eisenlohr, D. Jiang, M. Torrent-Moreno, L. Delgrossi, and H. Hartenstein, "Overhaul of ieee 802.11 modeling and simulation in ns-2," Proceedings of the 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems, ACM, 2007, pp. 159-168.

[6] M. A. Youssef, A. Vasan and R.E. Miller, Specification and Analysis of the DCF and PCF Protocols in the 802.11 Standard Using Systems of Communicating Machines, Proceedings of the 10 th IEEE International Conference on Network Protocols (ICNP'02) 1092-1648, 2002.

[7] L. Romdhani, Q. Ni, and T. Turletti, "Adaptive EDCF: enhanced service differentiation for IEEE 802.11 802.11 wireless ad-hoc networks," *Proc. IEEE WCNC'03*, pp. 1373-1378.

[8] N.S. Nandiraju, H. Gossain, D. Cavalcanti, K.R. Chowdhury, and D.P. Agrawal, "Achieving Fairness in Wireless LANs by Enhanced 802.11 DCF," the Proc. of IEEE Wireless and Mobile Computing, Networking and Communications (WiMob 2006), 2005.

 ISBN: 978-1-61208-203-5

Reduced Complexity Algorithm for Quantized Equal Gain Transmission Codebook over Closed Loop MIMO Systems

Noe Yoon Park, Xun Li, and Young Ju Kim
College of Electrical and Computer Engineering
Chungbuk National University, Chungbuk 361-763 Rep. of Korea
Email: {nypark, lixun, yjkim}@cbnu.ac.kr

***Abstract*—In this paper, a novel index search algorithm is proposed for quantized equal gain transmission (QEGT) codebook based closed-loop multiple-input multiple-output systems. The codewords are grouped into a small number of categories by utilizing Grassmannian beamforming criterion. Then, the optimal centroid is selected within those of the groups, which maximizes the signal-to-noise ratio or capacity criterion. Hereby, the optimal index of QEGT codebook is determined within the group of the selected centroid. Monte-Carlo simulations are presented to show that the codebook index search complexity is halved, whilst maintaining almost the same throughput.**

***Keywords-Codebook; Closed-loop MIMO; Limited feedback; Grouping; Grassmannian beamforming*.**

I. INTRODUCTION

For closed-loop multiple-input multiple-output (CL-MIMO) communication systems, the transmit beamforming (TxBF) technique alleviates the negative effect of fading channel by exploiting spatial diversity due to the increased number of MIMO fading channel paths. The TxBF requires feedback of channel state information (CSI) from the receiver to the transmitter. Such CSI feedback can potentially incur excessive overhead due to the multiplicity of channel coefficients, and thus a small number of feedback bits are sent via a feedback path for the transmitter to recreate the TxBF vector. These systems are known as limited feedback systems (see [1] and the references therein). To reduce the bandwidth requirement of the feedback systems, finite rate techniques have been proposed for the cases of TxBF [2]-[4]. In these limited feedback systems, the receiver chooses a precoding matrix from a finite set of precoding matrices, called codebook, on the basis maximizing the effective capacity or signal-to-noise (SNR) after combining, and sends the corresponding bits, which denotes index, to the transmitter. The codebook design strategies which have been suggested use either numerical optimization methods [4]-[8] or random vector quantization (RVQ) method [9]. Such random codebooks have been shown to be asymptotically optimal as the number of feedback bits and transmitted antennas increase [10], [11].

Among various codebooks for the TxBF, the QEGT codebook is the optimal precoding matrix for maximizing the capacity with a per-antenna equal power constraint. Also, it has modest transmit amplifier requirements than other TxBF techniques, since it does not require the antenna amplifiers to modify the amplitudes of the transmitted input signals [5]. Similar to other codebook based CL-MIMO systems, QEGT codebook needs a larger size codebook which gives better performance than a small size one, as the number of transmit antennas increases. In other words, the codebook size increases exponentially with the number of transmit antennas to maintain a given effective capacity or SNR loss with respect to the ideal non-quantized system [3], [4], [8]. The increased codebook size can cause a feedback delay due to the large amount calculation of excessive search (full index search) [4],[5], and also reduce the effectiveness of the precoding matrix at the transmitter [12]. It can be seen that the feedback delay may lead to negative effects on information capacity or symbol error rate [13]-[17]. Non-exhaustive methods for searching unstructured codebooks at the expense of increased memory requirements have been well researched in [18]. Hence, employing an efficient codebook index search algorithm becomes essential.

In this paper, an efficient codebook index search algorithm with Grassmannian beamforming criterion is proposed for finding optimal precoding matrix of QEGT codebook based CL-MIMO systems. Monte-Carlo simulations are presented to show that the normalized complexity is halved, whilst maintaining almost the same achievable throughput comparing to the full index search algorithm when the number of transmit antennas is more than three.

The remainder of the paper is organized as follows. Section II reviews CL-MIMO communication with TxBF. In Section III, we propose reduced complexity QEGT codebook index search algorithm which relies on a grouping strategy. In Section IV, the results of previous sections are demonstrated, and related discussions are given. Finally, concluding remarks are given in Section V.

II. SYSTEM OVERVIEW

The CL-MIMO system relying on the TxBF and using N_t transmit as well as N_r receive antennas is considered. The M-dimensional complex transmit symbol vector at the

 ISBN: 978-1-61208-203-5

channel instant m (for $m = 0, 1, \ldots$) is denoted by [1]$\mathbf{s}_m = [\mathbf{s}_{m,1} \cdots \mathbf{s}_{m,M}]^T \in \mathbb{C}^{M\times 1}$ with $\mathbf{s}_m \sim \mathcal{CN}(\mathbf{0}_{M\times 1}, \mathbf{I}_M)$ and $M \leq \min\{N_t, N_r\}$. The vector $\mathbf{W}\mathbf{s}_m$ is sent through the channel where $\mathbf{W} \in \mathbb{C}^{N_t\times M}$ is the precoding matrix. Then, the received signal vector $\mathbf{y}_m$ at the N_r receive antennas can be written as

$$\mathbf{y}_m = \sqrt{\frac{\rho}{M}}\mathbf{H}_m\mathbf{W}\mathbf{s}_m + \mathbf{n}_m, \quad (1)$$

where $\mathbf{n}_m \in \mathbb{C}^{N_r\times 1}$ denotes the noise vector with $\mathbf{n}_m \sim \mathcal{CN}(\mathbf{0}_{N_r\times 1}, \mathbf{I}_{N_r})$ and ρ denotes the SNR. The matrix $\mathbf{H}_m \in \mathbb{C}^{N_r\times N_t}$ represents uncorrelated Rayleigh falt fading channel matrix with i.i.d. entries distributed according to $\mathcal{CN}(0, 1)$.

The evolution of $\mathbf{H}_m$ is modeled by a first-order Gauss-Markov process [19]

$$\mathbf{H}_m = \epsilon\mathbf{H}_{m-1} + \sqrt{1-\epsilon^2}\mathbf{G}_m, \quad (2)$$

where $\mathbf{G}_m \in \mathbb{C}^{N_r\times N_t}$ has i.i.d. entries with distribution $\sim \mathcal{CN}(0, 1)$. The noise process $\mathbf{n}_m$ in (1) is independent of $\mathbf{G}_m$ and $\mathbf{H}_0$. The time correlation coefficient $\epsilon\,(0 \leq \epsilon \leq 1)$ represents the correlation between elements $h_{m,i,j}$ and $h_{m-1,i,j}$ (where $h_{m,i,j}$ denotes the (i, j) entry of $\mathbf{H}_m$). We assume all the elements of $\mathbf{H}_m$ have the same ϵ. The evolution variable ϵ obeys Jakes' model [20] according to $\epsilon = J_0(2\pi f_D T)$, where $J_0(\cdot)$ is the zeroth order Bessel function, T denotes the channel instantiation interval, and $f_D = \frac{v f_c}{c}$ denotes the maximum Doppler frequency using terminal velocity v, carrier frequency f_c, and $c = 3 \times 10^8 m/s$.

In a TxBF system, the key question is how to design $\mathbf{W}$ to maximize the system performance. For this reason, $\mathbf{W}$ should be chosen to maximize the receive SNR in order to minimize the average probability of error and maximize the capacity. In general, it can be determined by applying the singular vector decomposition (SVD) [21]. However, accurate quantization and feedback of this precoding matrix can require a large number of feedback bits quantized TxBF techniques provide a solution for this problem by quantizing the optimal precoder at the receiver. Specifically, the precoder is constrained to be one of N matrices, which is called a codebook. If the codebook of N matrices is known to both the transmitter and the receiver, $L = log_2 N$ bits of feedback are required for indicating the index of the appropriate precoding matrix [22]. Denote the precoding QEGT codebook $\mathcal{W} = \{\mathbf{W}_k\}_{k=1}^N$ and $\mathbf{W}_k \in \mathcal{U}(N_t, M)$. A procedure to generate the QEGT codebook for TxBF system is clearly given in six steps [5].

[1]a bold lower case letter $\mathbf{a}$ denotes the vector, a bold capital letter $\mathbf{A}$ denotes a matrix, $\mathbf{A} \in \mathbb{C}^{m\times n}$ denotes complex matrix $\mathbf{A}$ having m row and n column, $\mathbf{A}^H$ denotes the conjugation transposition of matrix $\mathbf{A}$, $\mathbf{I}_M$ denotes the $M \times M$ identity matrix, $\mathbf{0}_{m\times n}$ denotes the $m \times n$ zero matrix, $\mathcal{U}(m, n)$ denotes the set of $m \times n$ matrices with orthogonal columns, $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, $\log_2(\cdot)$ denotes the base two logarithm, and $\det(\cdot)$denotes the determinant of a matrix.

We assume that the receiver perfectly knows the current CSI by channel estimation algorithms. In order to maximize the system performance, two optimum codebook index selection criterions are presented: one is the SNR selection criterion and another is the capacity selection criterion. In the SNR selection criterion, the $\mathbf{W}$ should be selected to maximize the receive SNR, thus the selection criterion can be expressed as follows [4]

$$\mathbf{W} = \arg\max_{\mathbf{W}_k\in\mathcal{W}} \|\mathbf{H}_m\mathbf{W}_k\|_F. \quad (3)$$

Another precoding matrix selection criterion is the maximum capacity, which can be written as follows [23]

$$\mathbf{W} = \arg\max_{\mathbf{W}_k\in\mathcal{W}} \log_2 \det\left(\mathbf{I}_M + \frac{\rho}{M}\mathbf{W}_k^H\mathbf{H}_m^H\mathbf{H}_m\mathbf{W}_k\right). \quad (4)$$

From (3), (4), the optimal precoding matrix $\mathbf{W}$ satisfy both SNR and capacity maximization criterions.

III. Proposed Index Search Algorithm

In order to reduce the complexity of index search algorithm for QEGT, we propose a new codeword grouping algorithm. In the second subsection, the optimal precoding matrix index selection algorithm was presented among the generated groups.

A. Grouping the codewords of QEGT

The prcoding matrices in the codebook are divided into a given number groups. Assuming that the QEGT codebook elements are uniformly quantized, the precoding matrices can be arranged into P groups, each group having Q precoding matrices $(N = P \times Q)$. The proposed grouping strategy uses Grassmannian beamforming criterion to generate P groups of codeword. Also, another grouping strategy using Lloyd's algorithm will be addressed briefly to compare performance for the proposed algorithm, which has already studied in [25]. The major difference between these two grouping strategies is that the key idea of proposed algorithm uses index elimination, while the Voronoi cell is used in [25]. Also, the proposed algorithm has characteristic that the precoding matrices in the QEGT codebook are distributed always evenly to each group, while the grouping strategy using Lloyd's algorithm is not always doing so. $i.e.$, if we have N = 16, P = 4, Q for the proposed algorithm is always 4, Q for Lloyd's algorithm has a value between 3 and 5. Thus, the calculation of proposed algorithm is always minimum comparing to the approach in [25]. This can be extended more generally, which will be shown in Section IV.

1) Grassmannian beamforming criterion: Using Grassmannian beamforming criterion, it can be maximized that the minimum distance between any pair of lines spanned by the codebook matrices on Grassmann manifold [4], [24]. And it provides an approach for finding optimal line packings. We outline the process as the following.

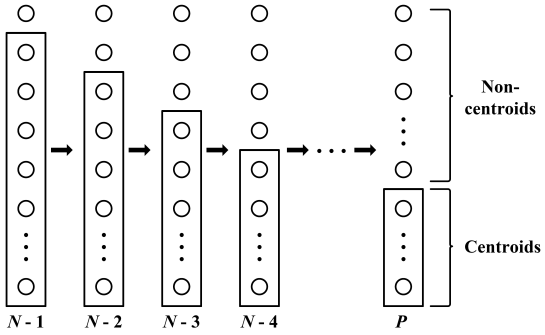


Figure 1. Centroid selection using Grassmannian beamforming criterion.

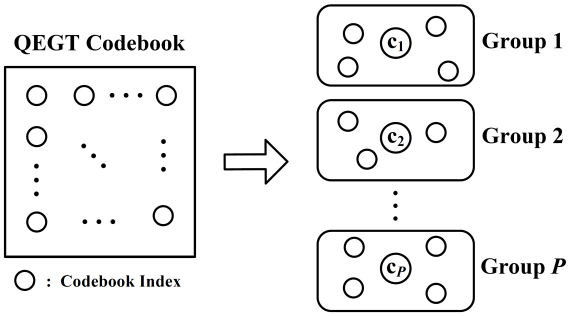


Figure 2. Grouping the precoding matrix using chordal distances.

- Step 1: Generate candidate reference codebooks $\{\mathcal{W}_i^c\}_{i=1}^N$ by deleting the i-th index matrix from $\mathcal{W}$. Thereby the number of candidate reference codebooks is $N = card\,(\mathcal{W})$, where $card(\cdot)$ denotes the cardinality of a set. Let $N^c = card\,(\mathcal{W}_i^c)$.
- Step 2: Select the optimal candidate codebook $\mathcal{W}_{opt}^c$ from the candidate codebooks which has maximized minimum distance between each pair of matrices. This can be expressed as follows

$$d\,(\mathbf{W}_k, \mathbf{W}_l) = \sqrt{1 - \left|\mathbf{W}_k^H \mathbf{W}_l\right|^2}, \tag{5}$$

$$\delta\,(\mathcal{W}_i^c) = \min_{(\mathbf{W}_k, \mathbf{W}_l) \in \mathcal{W}_i^c} d\,(\mathbf{W}_k, \mathbf{W}_l)\,, \tag{6}$$

and

$$\mathcal{W}_{opt}^c = \underset{\mathcal{W}_i^c \in \{\mathcal{W}_i^c\}_{i=1}^N}{\arg\max}\ \delta\,(\mathcal{W}_i^c)\,. \tag{7}$$

- Step 3: If N^c is not equal to P, let $\mathcal{W} = \mathcal{W}_{opt}^c$ which is previously selected. And repeat step 1 and step 2 until $N^c = P$, as shown in Fig. 1.
- Step 4: The deleted precoding matrices (in step 1) are assigned to its nearest centroid by using the chordal distance as in (5). This procedure generates new groups which only have one centroid in each of them, as shown in Fig. 2, where c_p, $p = 1,\ 2,\ \ldots,\ P$ are the centroid indices.

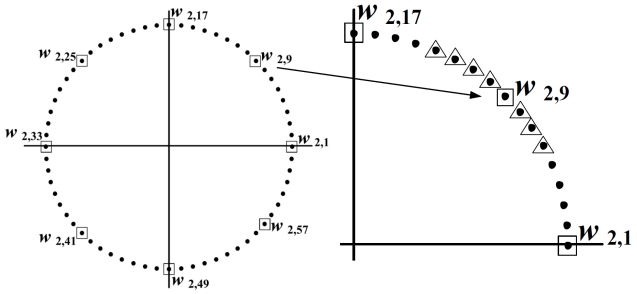


Figure 3. Example of QEGT codebook grouping when N_t=2, N=64 and M=1.

2) Lloyd's algorithm: To illustrate the grouping strategy with Lloyd's algorithm, we consider the case of N_t=2, N=64 and M=1. The 1st element of the TxBF vectors corresponding to the 1st transmit antenna can be forced to be real-valued owing to the rotation-invariant property of BF vectors [4]. Then, the 2nd element of the TxBF vectors corresponding to the 2nd transmit antenna can be plotted on the complex-valued plane, as seen in Fig. 3, which illustrates eight groups of TxBF vectors as the result of Lloyds clustering algorithm. In the notation of $\mathbf{w}_{i,j}$ seen in Fig. 3, i represents the i-th element of the TxBF vector corresponding to the i-th transmit antenna and j denotes the index of the TxBF vector in the QEGT codebook. Fig. 3 also magnifies a group of eight TxBF vectors at the right, where the cluster centroid is indicated by a square mark ($\square$), while the others are indicated by triangles ($\triangle$). In the same way, we can group the precoding matrices in the QEGT codebook, when the number of transmit antennas is higher than or equal to three. The more detailed grouping strategy using Lloyd's algorithm has been well documented in [25].

B. Group index selection criterion for QEGT codebook

Among the various centroids, the best of the centroids c_{opt} is selected, which maximized the receive SNR

$$c_{opt} = \underset{c_p \in C_{set}}{\arg\max} \left\|\mathbf{H}_m \mathbf{W}_{c_p}\right\|_F, \tag{8}$$

or channel capacity

$$c_{opt} = \underset{c_p \in C_{set}}{\arg\max} \log_2 \det\left(\mathbf{I}_M + \frac{\rho}{M} \mathbf{W}_{c_p}^H \mathbf{H}_m^H \mathbf{H}_m \mathbf{W}_{c_p}\right), \tag{9}$$

where C_{set} is the set of $\{c_1,\ c_2,\ \ldots,\ c_p\}$. And then, the optimal precoding matrix index m_{opt} is determined within the group of the selected best centroid, which maximized the receive SNR as follows

$$m_{opt} = \underset{m_q \in M_{set}}{\arg\max} \left\|\mathbf{H}_m \mathbf{W}_{m_q}\right\|_F, \tag{10}$$

or channel capacity

$$m_{opt} = \underset{m_q \in M_{set}}{\arg\max} \log_2 \det\left(\mathbf{I}_M + \frac{\rho}{M} \mathbf{W}_{m_q}^H \mathbf{H}_m^H \mathbf{H}_m \mathbf{W}_{m_q}\right), \tag{11}$$

 ISBN: 978-1-61208-203-5

where m_q, $(q = 1,\ 2,\ \ldots,\ Q)$ represents the precoding matrices of the chosen centroid group and M_{set} is the set of $\{m_1,\ m_2,\ \ldots,\ m_q\}$.

From the above process, proposed scheme needs only the search of $P+Q$ indices within QEGT codebook. It is much smaller than the number of full index search algorithm in (3), (4). For this reason, the entire search complexity is significantly reduced. Also, considering the highly temporally correlated fading channels, the group of the selected centroid as in (8), (9) almost constants. Therefore the entire search complexity is more reduced by researching only the Q indices.

The problem now becomes how to determine the number of groups, P and the number of elements Q in the group, when QEGT codebook size N is fixed. The proposed grouping rule can be applied that the number of groups, P should be the same or the nearest integer to the value of N/P when the precoding matrix in the QEGT codebook are distributed almost evenly to each group.

IV. Simulation Results

In this section, we perform Monte-Carlo simulations to investigate the achievable throughput performance of the proposed algorithm in MIMO channels. The CL-MIMO system is equipped with N_t transmit antennas, and N_r receive antennas, we use the notation $N_t \times N_r$ to denote N_t transmit and N_r receive antenna system. In order to show the performance in vehicular environments, simulation parameters employed in IEEE 802.11p/WAVE standard [26] are used. In IEEE 802.11p/WAVE, the TxBF system assumes mobile speed 50km/h, carrier frequency of 5.9GHz, bandwidth of 10MHz, and feedback interval of 5ms, thus, typical time correlation coefficient is ϵ=0.020. In our systems, only low mobile speed is supported (less than 60km/h), this is supported in 3GPP LTE and IEEE 802.16p standards. And adaptive modulation coding is assumed for orthogonal frequency division multiplexing system whose DFT/IDFT size is 512. The simulations were run with over 1.5 million iterations per SNR point. Codebook for the QEGT system was designed based on Grassmannian beamforming criterion [5]. The entire codebook size for 3 transmit antennas is 16, and the entire codebook size for 4 and 5 transmit antennas is 64. The maximum capacity selection criterion is used for the optimal precoding matrix selection. The feedback channel is assumed to be error free. Also, we assumed that the channel estimation and synchronization are perfect and there is no spatial correlation amongst transmit and receive antennas.

The achievable throughput performance of the proposed and the full index search ones which have 3×2, 4×3 and 5×4 systems are shown in Fig. 4. We assume P is 4 and 8 for 3 transmit antennas and 4 or 5 transmit antennas, respectively. Fig. 4 shows the performance of proposed algorithm, the proposed algorithm (curve label 'proposed algorithm') performs almost the same as those

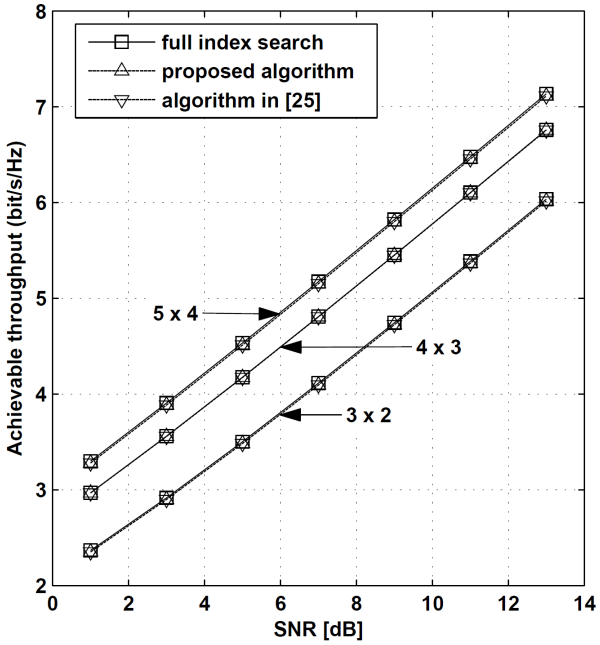


Figure 4. Throughput versus SNR when mobile speed is 50km/h.

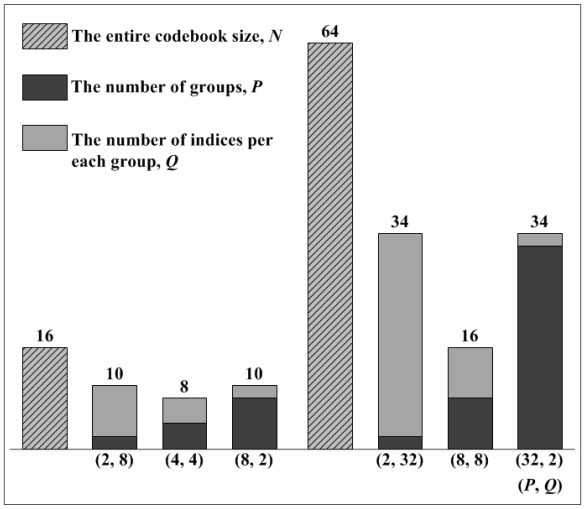


Figure 5. Comparison of the calculated indices for QEGT codebook both the full index search algorithm and the proposed algorithm.

of the approach in [25] (curve label 'algorithm in [25]'). And those performance almost achieve the full index search algorithm (curve label 'full index search').

Fig. 5 shows how much the QEGT codebook index search complexity is reduced in comparison to the full index search algorithm. When the proposed codebook index search algorithm using group strategy based on Grassmannian beamforming criterion is applied, $i.e.$, we have $(N,\ P,\ Q) = (16,\ 4,\ 4)$ or $(64,\ 8,\ 8)$, we have almost the same performance comparing to the full index search ones, while maintaining the lowest codebook index search

 ISBN: 978-1-61208-203-5

complexity. From the results of Fig. 5, the optimal grouping is that the value of P should be nearly the same at the value of Q. In other words, the value of Q may vary around the nearest integer of N/P. Thereby, the codebook index search complexity is halved, whilst maintaining almost the same throughput when the number of transmit antennas is more than three.

V. Conclusion

In this paper, we have investigated the codebook index search problem for CL-MIMO systems. A complexity reduced index search algorithm for QEGT codebook is proposed which uses grouping strategy based on Grassmannian beamforming criterion. From the simulation results, as the QEGT codebook size increases, the QEGT codebook index search complexity of the proposed algorithms were significantly decreased comparing with that of the full index search algorithm. Moreover, the achievable throughput performance of our proposed algorithm were almost the same as those of the existing full index approaches.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2012R1A1B6002111).

References

[1] D. J. Love, R. W. Heath, Jr., W. Santipach and M. L. Honig "What is the value of limited feedback for MIMO channels?," *IEEE Commun. Magazine*, vol. 42, pp. 54-59, Oct. 2004.

[2] A. Narula, M. J. Lopez, M. D. Trott, and G. W. Wornell, "Efficient use of side information in multiple-antenna data transmission over fading channels," *IEEE J. Selected Areas of Commun.*, vol. 16, pp. 1423-1436, Oct. 1998.

[3] K. K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna systems," *IEEE Trans. on Information Theory*, vol. 49, pp. 2562-2579, Oct. 2003.

[4] D. J. Love, R. W. Heath, Jr., and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Trans. on Information Theory*, vol. 49, pp. 2735-2747, Oct. 2003.

[5] D. J. Love and R. W. Heath, Jr., "Equal Gain Transmission in Multiple-Input Multiple-Output Wireless Systems," *IEEE Trans. on Communications*, vol. 51, pp. 1102-1110, July 2003.

[6] C. R. Murthy and B. D. Rao, "Quantization methods for equal gain transmission with finite rate feedback," *IEEE Trans. on Signal Processing*, vol. 55, pp. 233-245, Jan. 2007.

[7] J. C. Roh and B. D. Rao, "Transmit beamforming in multiple-antenna systems with finite rate feedback: a VQ-based approach," *IEEE Trans. on Information Theory*, vol. 52, pp. 1101-1112, Mar. 2006.

[8] P. Xia, S. Zhou and G. B. Giannakis, "Achieving the Welch bound with difference sets," *IEEE Trans. on Information Theory*, vol. 51, pp. 1900-1907, May 2005.

[9] W. Santipach and M. L. Honig, "Signature optimization for CDMA with limited feedback," *IEEE Trans. on Information Theory*, vol. 51, pp. 3475-3492, Oct. 2005.

[10] W. Dai, Y. Liu, and B. Rider, "Quantization Bounds on Grassmann Manifolds and Applications to MIMO Communications," *IEEE Trans. on Information Theory*, vol. 54, pp. 1108-1123, Mar. 2008.

[11] W. Santipach and M. L. Honig, "Capacity of a multiple-antenna fading channel with a quantized precoding matrix," *IEEE Trans. on Information Theory*, vol. 55, pp. 1218-1234, Mar. 2009.

[12] K. Huang, B. Mondal, R. W. Heath, Jr., and J. G. Andrews, "Effect of feedback delay on multi-antenna limited feedback for temporallycorrelated channels," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Francisco, CA, USA, Nov. 2006.

[13] J. Du, Y. Li, D. Gu, A. F. Molisch, and J. Zhang, "Estimation of performance loss due to delay in channel feedback in MIMO systems," in *Proc. IEEE Veh. Technology Conf.*, vol. 3, pp. 1619-1622, Sep. 2004.

[14] E. Au, S. Jin, M. R. McKay, W. Mow, X. Gao, and I. B. Collings, "Analytical performance of MIMO-SVD systems in Ricean fading channels with channel estimation error and feedback delay," *IEEE Trans. Commun.*, vol. 7, no. 4, pp. 1315-1325, Apr. 2008.

[15] Y. Isukapalli and B. D. Rao, "Finite rate feedback for spatially and temporally correlated MISO channels in the presence of estimation errors and feedback delay," in *Proc. IEEE GLOBECOM*, pp. 2791-2795, Nov. 2007.

[16] K. Kobayashi, T. Ohtsuki, and T. Kaneko, "MIMO systems in the presence of feedback delay," in *Proc. IEEE Int. Conf. Communications*, vol. 9, pp. 4102-4106, Jun. 2006.

[17] S. H. Ting, K. Sakaguchi, and K. Araki, "A Markov Kronecker model for analysis of closed-loop MIMO systems," *IEEE Commun. Lett.*, vol. 10, pp. 617-619, Aug. 2006.

[18] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. on Information Theory*, vol. 44, pp. 2325-2383, Oct. 1998.

[19] R. H. Etkin and D. N. C. Tse, "Degree of freedom in some underspread MIMO fading channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1576-1608, Apr. 2006.

[20] J. G. Proakis, *Digital Communications*, 4th edition, McGraw Hill, 2000.

[21] D. J. Love and R. W. Heath, Jr., "Limited feedback unitary precoding for orthogonal space-time block codes," *IEEE Trans. on Signal Processing*, vol. 53, pp. 64-73, Jan. 2005.

[22] J. G. Andrews, A. Ghosh, and R. Muhamed, Fundamentals of WiMAX understanding broadband wireless networking, *Prentice Hall*, 2007.

 ISBN: 978-1-61208-203-5

[23] T. J. Kim, D. J. Love and B. Clerckx, “MIMO system with limited rate differential feedback in slowly varying channels,” *IEEE Trans. on Commun.*, vol. 59, no. 4, pp. 1175-1189, Apr. 2011.

[24] A. Bang and D. Y. Nogin, “Bounds on packings of spheres in the Grassmann manifold,” *IEEE Trans. on Information Theory*, vol. 48, pp. 2450-2454, Sep. 2002.

[25] Y. J. Kim, S. H. Won, N. Y. Park, and L. Hanjo, “Reduced-complexity transmit-beamforming codebook search algorithm,” *Elec. Letter*, vol. 47, pp. 938-939, Aug. 2011.

[26] IEEE Std 802.11p, “Wireless LAN medium access control and physical layer specifications: Wireless access in vehicular environments,” July 2010.

 ISBN: 978-1-61208-203-5

Performance Improvement of Differential Codebooks with Noisy Feedback Channels

Xun Li, NoeYoon Park and YoungJu Kim
College of Electrical and Computer Engineering
Chungbuk National University
Chungbuk 361-763 Rep. Korea
Email: {lixun, nypark, yjkim}@cbnu.ac.kr

***Abstract*—In this paper, a differential codebook indexing scheme is proposed for limited feedback system over noisy feedback channels. A lot of research has been done for limited feedback system assuming error free feedback. In practical systems, the feedback information experiences noisy feedback channels which cause feedback information partially all totally useless. Prior research about differential precoding focuses on the codebook design criterion which minimize the quantization distortion. The proposed scheme focuses on how to minimize the effect due to feedback errors. The relationship of feedback errors and limited feedback system performance is analyzed in this paper. Using the analytical results, an optimal differential codebook indexing scheme is proposed to improve the system performances when the feedback bits less than 3 bits and exceed 4 bits, respectively. From some selected numerical results, the proposed differential codebook indexing scheme provides non-negligible performance improvements in terms of average bit error rate than the systems without indexing.**

Keywords-Indexing, Differential codebook, Temporal correlation, Limited feedback.

I. INTRODUCTION

Transmit beamforming for multiple-input multiple-output (MIMO), which is also known as precoding, has been widely adopted in wireless communication standards (WiMAX, 3GPP-LTE [1]). It uses some type of quantized channel state information (CSI) at the transmitter to offer good trade-off between performance gain and the required amount of feedback bits [2]–[4]. The accuracy of CSI at the transmitter depends on the feedback bits used. For block to block fading channel model, the channel realization is considered to change independently. But in low mobility scenarios, the temporal correlation always existed between adjacent channel realizations. Quantized differential feedback improves the quantization resolution utilizing the temporal correlation of the channels [5], [6]. In temporally correlated channels, the channel realization is changed slowly, as well as the optimal precoder. Thereby, quantizing the whole channel space is waste the feedback resource. Quantized differential feedback scheme quantizes the specific channel subspace instead of whole space, and the codebook for current time instant is various over time and depends on the previous precoder and the channel long term statistic [7]–[9]. The proposed schemes indicate that quantized differential feedback scheme can greatly improve the system performance.

In prior researches, the feedback channel is assumed to be error free and delay free for simplicity. In this case, the indexes are arbitrarily assigned to the set of codewords. However, the feedback error cannot be avoided although many techniques (lower modulation order, high channel coding redundancy, etc.) are used in feedback transmission [1]. The feedback errors causes that the transmitter applies precoding with undesired precoder. The effects of feedback error to the performance of general codebook based precoding system have been analyzed, and the principle of codebook index algorithms have been proposed in [10], [11]. And index assignment scheme for beamforming system are proposed to minimize the effects from feedback errors. These algorithms demonstrated the procedures of codebook index which have no consideration of computation complexity.

In this paper, a complex reduced codebook index algorithm is proposed when the feedback information is more than 4 bits. The proposed algorithm can be realized with low complexity circuit. Also, the index assignment scheme is applied in quantized differential feedback system. The differential precoding system with proposed codebook index algorithms effectively lowers the error floor introduced by the feedback errors. We analyze the effects of feedback errors to the limited feedback system, the performance of proposed scheme is evaluated and compared with the performance of the long term evolution (LTE) codebook with or without noisy feedback channels, respectively. In order to make the index assignment is applicable when the number of feedback bits is more than 4 bits, a suboptimal index assignment scheme is proposed which shows more flexible trade-off between performance and calculation complexity. Without loss of the generality, we compared the performance of the LTE codebook with the differential codebook proposed in [8], which preserves per-antenna equal power constraint property like the former codebook.

The rest of this paper is organized as follows: Section II shows the system overview of general limited feedback system and quantized differential feedback system. Section III introduces the index assignment schemes in codebook based limited feedback systems. Section IV illustrates the application of index assignment scheme to the quantized differential feedback system. And the simulation results are shown in this part. Finally, the conclusions are shown in

 ISBN: 978-1-61208-203-5

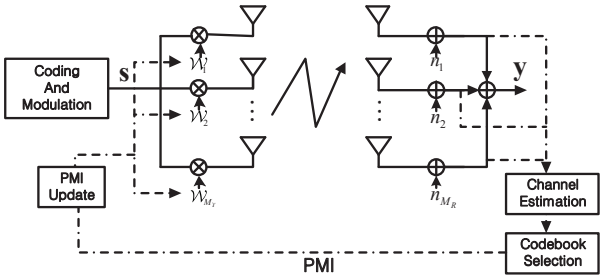


Figure 1. Block diagram of limited feedback system.

Section V.

II. System Overview

In this section, the limited feedback system and quantized differential feedback system are introduced. The codebook design criterion is also introduced.

A. Limited Feedback System

A MIMO system employing M_t transmit antennas and M_r receive antennas is assumed in this paper. The block diagram is shown in Fig. 1. The transmit symbols at the time instant τ are denoted by $\mathbf{s}_\tau = [s_{\tau,1}, \cdots, s_{\tau,V}]^T$, where V denotes the number of data streams (also called transmission rank), and $1 \leq V \leq min\{M_t, M_r\}$. The received signal is represented by

$$\mathbf{y}_\tau = \sqrt{\frac{\rho}{V}}\mathbf{H}_\tau\mathbf{F}_\tau\mathbf{s}_\tau + \mathbf{n}_\tau, \tag{1}$$

where ρ is signal-to-noise ratio (SNR), $\mathbf{F}_\tau \in \mathbb{C}^{M_t \times V}$ denotes the precoder at time instant τ. Without loss of generality, we assume that each column of $\mathbf{F}$ is normalized. $\mathbf{n}_\tau$ denotes the additive white Gaussian noise (AWGN) vector at time instant τ with distribution of $\mathcal{CN}(0,1)$. The matrix $\mathbf{H}_\tau \in \mathbb{C}^{M_r \times M_t}$ represents a spatially uncorrelated but temporally correlated Rayleigh fading channel, which is modeled by the first-order Gauss-Markov process

$$\mathbf{H}_\tau = \epsilon\mathbf{H}_{\tau-1} + \sqrt{1-\epsilon^2}\mathbf{G}_\tau, \tag{2}$$

where $\mathbf{G}_\tau$ has the same size of $\mathbf{H}_\tau$ with i.i.d entries and represents the evolution of $\mathbf{H}_\tau$. The $\epsilon \in [0,1]$ denotes the time correlation between the channel coefficient of adjacent time instants. In this paper, the ϵ obeys Jakes' model [12], [13].

Assuming perfect channel knowledge of the current channels at the receiver, the mutual information is known to be

$$I(\mathbf{F}_\tau) = \log_2\left(\det\left(\mathbf{I}_V + \frac{\rho}{V}\mathbf{F}_\tau^*\mathbf{H}_\tau^*\mathbf{H}_\tau\mathbf{F}_\tau\right)\right). \tag{3}$$

The optimal precoder without quantization can be obtained via singular value decomposition (SVD) of channel. In limited feedback systems, a codebook is known by the transmitter and receiver, the precoder $\mathbf{F}_\tau$ is selected from the codebook $\mathcal{F}_\tau = \{\mathbf{F}_{\tau,i}\}_{i=1}^N$, where N denotes the codebook size. The receiver selects favorite precoder from the codebook and sends the index back to transmitter. There is no argument on that precoder selection is based on mutual information maximization criterion, which is shown as the following

$$\mathbf{F}_\tau = \underset{\mathbf{F}_{\tau,i}\in\mathcal{F}_\tau}{\arg\max}\left\{I(\mathbf{F}_{\tau,i})\right\}. \tag{4}$$

B. Differential Feedback Framework

In temporal correlated channels, the quantized differential feedback can virtually increase the codebook size. Quantizing the specific subspace instead of the whole channel space improves the quantization resolution. The differential codebook generates points on Grassmann manifold which are centered by the previous precoder. The differential codebook is shared by the transmitter and receiver, as well as the codebook update criterion. A quasi-diagonal differential codebook is proposed in [5]. The spherical cap differential codebook with adaptive cap radius is proposed in [7]. These differential codebooks can be categorized into total power constraint codebook since they change the power and phase of each antenna to achieve the maximum throughput. In LTE standard and its advanced version (LTE-A), the equal gain transmission property is considered to be the basic requirement of the precoding scheme. In order to fairly compare the performance degradation of limited feedback system and quantized differential feedback system over noisy feedback channels, we use the differential equal gain transmission (DEGT) codebook proposed in [8].

The DEGT has flexible trade-off between performance and codebook design complexity. Assuming the scalar design scheme which utilizes the structure of initial codebook, the DEGT codebook $\mathcal{F}' = \{\mathbf{F}'_i\}_{i=1}^N$ can be designed as the following

$$\mathbf{F}'_i = e^{j\alpha\angle\mathbf{F}_{i,0}} \tag{5}$$

where $\exp(\cdot)$ denotes the exponential function, $j = \sqrt{-1}$, $\angle\mathbf{A}$ denotes the phase matrix of $\mathbf{A}$, and $\mathbf{F}_{i,0}$ denotes the i-th codeword of initial codebook $\mathcal{F}_0$. The scalar factor α decides the range of differential codebook covered the Grassmann manifold. It should be designed appropriately according to the channel temporal correlation and can be determined using iterative simulation. The capacity-α relationship for different channel temporal correlation is illustrated in Fig. 2. The codebook update criterion can be expressed as follows

$$\mathbf{F}_{i,\tau} = \mathbf{F}_{i,\tau-1} \circ \mathbf{F}'_i \tag{6}$$

where $\circ$ denotes the matrix element-wise multiplication. Note that, the codeword index in $\mathcal{F}_\tau$ selected by the receiver also is the differential codeword index in $\mathcal{F}'$. Thereby, the whole codebook $\mathcal{F}_\tau$ is not necessary to be constructed at the transmitter.

 ISBN: 978-1-61208-203-5

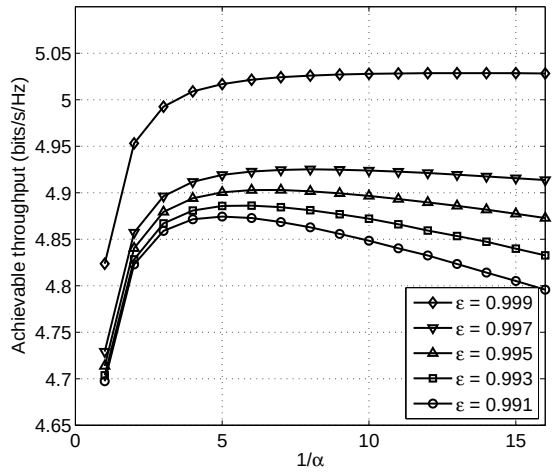


Figure 2. The optimal scalar factor value for different mobile speed.

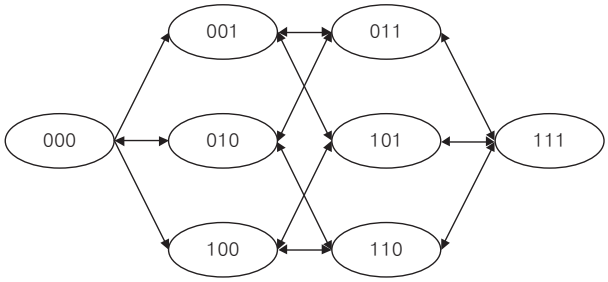


Figure 3. Example of 1 bit error flow chart (3 bits feedback).

III. Index Assignment for Differential Codebook

In this part, the codebook index assignment scheme is introduced. The codebook within 3 bits feedback can be calculated directly. For the codebook over 4 bits feedback, the computation complexity is too huge to calculate directly. We illustrate a grouping index assignment to reduce the complexity.

A. Codebook Index Assignment within 3 bits

The illustration of index variation when 1 bit error occurs in 3 bits feedback sequence is shown in Fig. 3. In this paper, we assume that 1 bit error occurs per feedback at most, since it already is a high probability. Based on the relationship shown in Fig. 3, the codebook index assignment procedure can be applied as follows

1) Generate all possible combinations of the N codewords in the codebook with different sort. The number of the possible combinations is $N!$.
2) Calculate the total Hamming distance between pair of different codewords of each combination. This is shown in follows

$$C = \sum_{i=0}^{N-2} \sum_{j=i+1}^{N-1} \Psi\left\{d\left(I_i, I_j\right)\right\} \left\| {\mathbf{F}'_i}^H \mathbf{F}'_j \right\|_F^2, \quad (7)$$

where I_i denotes the binary format of i-th codeword index. $d(x, y)$ denotes the Hamming distance between binary sequence x and y. The function $\Psi\cdot$ is shown as follows

$$\Psi\left\{a\right\} = \begin{cases} 1 & \text{if } a = 1 \\ 0 & \text{otherwise} \end{cases}, \quad (8)$$

The optimal codebook can be found by searching for the largest C.

B. Codebook Index Assignment over 4 bits

The number of combinations of the codebook with more than 4 bits feedback is too large to calculate by computer. We proposed a suboptimal codebook index assignment scheme by dividing the codewords into several groups called reference codebook. This shows flexible trade-off between calculation complexity and performance. Suppose the size of original codebook is N. The processes of reference codebook generation can be applied as follows

1) Generate $N = card\left(\mathcal{F}'\right)$ reference codebooks with $N - 1$ codewords in each reference codebook by deleting one codeword from the original codebook. The function $card(\cdot)$ denotes the cardinality of a set.
2) Find the codebook with maximized minimum Chordal distance between each pair of codewords in the reference codebook as shown in the following

$$\mathcal{W}_t = \underset{\mathcal{W}_k(1 \le k \le N)}{\arg\max} \left\{ \min_{\mathbf{F}'_i, \mathbf{F}'_j \in \mathcal{F}'_k} d\left(\mathbf{F}'_i, \mathbf{F}'_j\right) \right\}_{i=1}^{N}, \quad (9)$$

$$d\left(\mathbf{F}'_i, \mathbf{F}'_j\right) = \sqrt{M - \left\| \mathbf{F}'_i {\mathbf{F}'_j}^H \right\|_F^2} \quad (10)$$

3) Repeat the step 1) and 2) until the size of reference codebooks are R which is possible to use the algorithm introduced in Section III-A. The optimal reference codebook $\mathcal{W} = \{\mathbf{W}_i\}_{i=1}^{R}$ is obtained by sorting the reference codebook with that algorithm. We assume the set $\mathcal{W}' = \mathcal{F}' - \mathcal{W}$ which contains the deleted codewords $\mathcal{W}' = \{\mathbf{W}'_i\}_{i=1}^{N-R}$. We considered the optimal reference codebook has R groups and each group has single element at the first.
4) Move the codewords to the groups from $\mathcal{W}'$. The group index can be determined as the following

$$r = \underset{1 \le i \le R, 1 \le j \le N-R}{\arg\max} \left\{ \left\| {\mathbf{W}_i}^H \mathbf{W}'_j \right\|_F \right\}, \quad (11)$$

5) Repeat step 4) until all elements in $\mathcal{W}'$ is moved to $\mathcal{W}$. Sort the each group using the algorithm in Section III-A. Appending the index sequence of the codewords in group to the group index sequence. The suboptimal

 ISBN: 978-1-61208-203-5

codebook index assignment can be finished by these procedures.

IV. Simulation Results and Discussions

Monte-Carlo simulation is employed to obtain the performances of proposed scheme and conventional schemes. A MIMO system is considered which has 4 transmit antennas and 2 receive antennas. The number of data stream is one. The channel is modeled with first order Gaussian Markov process as described in Section II-A. The number of feedback bits is set to be 4 to consistent with LTE standard. The temporal correlation factor ϵ is assumed to be 0.991 and 0.997 which approximate to be 10 km/h and 3 km/h for LTE system. Correspondingly, the differential codebook scalar factor is set to be 5 and 8 respectively.

For quantized differential feedback system, once the error takes place in feedback, the codebook saved at transmitter and receiver becomes different and the difference will be accumulated. The initial codebook will be launched per 100 iterations in the simulation to break the accumulation.

Fig. 4 shows the comparison of LTE codebook and the differential codebook introduced in Section II-B. We assume the user equipment mobility is 10 km/h, and the scalar factor is 5. The bit error rate (BER) of feedback channel is assumed to be 10^{-3}.The performances are same for LTE codebook with or without index assignment. But the differential codebook shows performance improvement by using index assignment. Note that, the LTE codebook outperformed DEGT codebook after 9 dB since the error accumulation problem degrades the performance of DEGT scheme.

Figs. 5 and 6 show the comparison of DEGT codebook with or without index assignment when feedback error is 10^{-3} and 10^{-4}, respectively. The mobility of user equipment is 10 km/h in Fig. 5 and that is 3 km/h in Fig. 6. The index assignment scheme provides significant performance improvement in both low mobility and high mobility scenario.

Fig. 7 illustrates the relationship between the BER performance and initial codebook launching interval. The SNR is 10 dB, the mobility of user equipment is 10 km/h. The BER of feedback channel is assumed to be 10^{-3}. The index assignment scheme can increase the initial codebook launching interval.

V. Conclusions

In this paper, an index assignment algorithm for quantized differential feedback scheme is proposed. By using the index assignment to minimize the effects from the feedback error, the proposed scheme significantly improves the BER performance of the differential precoding system in the noisy feedback channels. For the codebook with more than 4 feedback bits, we also introduced a practical algorithm to make index assignment realizable.

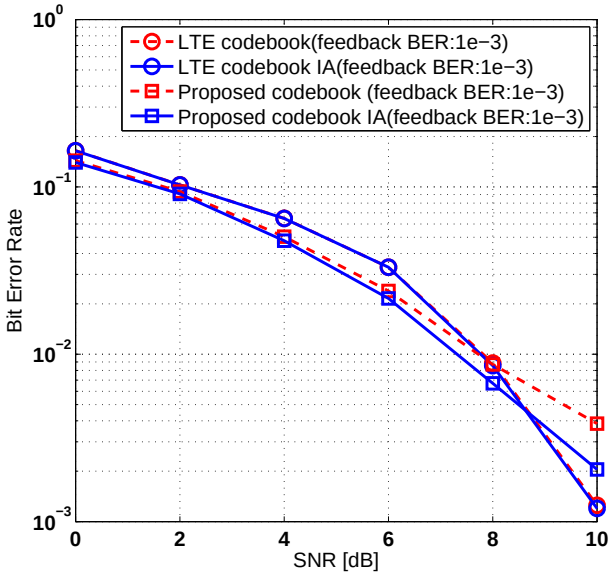


Figure 4. Performance comparison of differential codebook and LTE codebook.

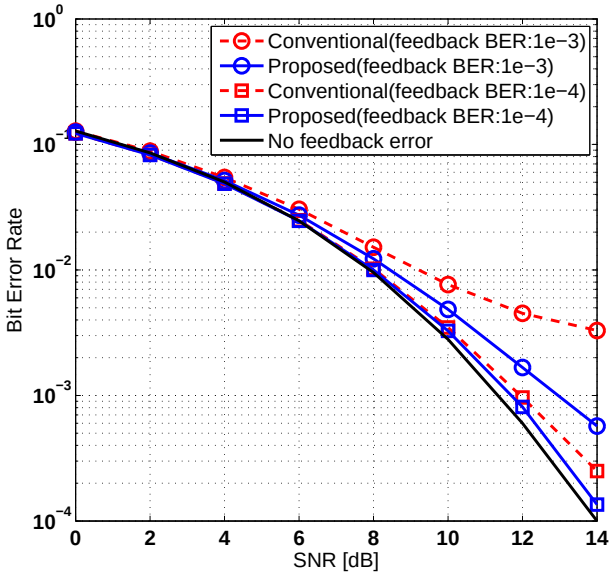


Figure 5. Performance comparison of index reassigned codebook ($f = 5$).

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012R1A1B6002111)

References

[1] *3GPP TS 36.211: Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation*, 3GPP Std.

 ISBN: 978-1-61208-203-5

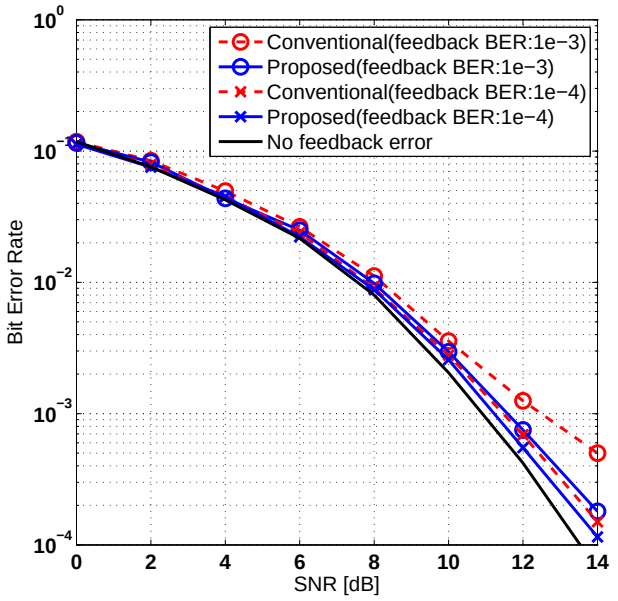


Figure 6. Performance comparison of index reassigned codebook ($f = 8$).

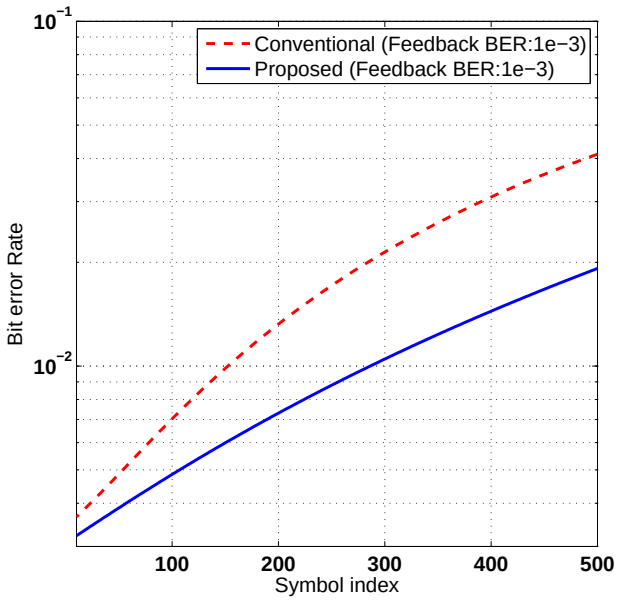


Figure 7. Performance comparison of retransmission timing of initial codebook ($f = 5$, SNR=10 dB).

[2] D. J. Love, J. Heath, R. W., and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2735–2747, 2003.

[3] D. J. Love and R. W. Heath, "Limited feedback unitary precoding for spatial multiplexing systems," *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2967–2976, 2005.

[4] S. A. Jafar and S. Srinivasa, "On the optimality of beamforming with quantized feedback," *IEEE Trans. Commun.*, vol. 55, no. 12, pp. 2288–2302, 2007.

[5] T. Abe and G. Bauch, "Differential codebook mimo precoding technique," in *Proc. IEEE Global Telecommunications Conf. GLOBECOM '07*, 2007, pp. 3963–3968.

[6] T. Kim, D. J. Love, B. Clerckx, and S. J. Kim, "Differential rotation feedback mimo system for temporally correlated channels," in *Proc. IEEE Global Telecommunications Conf. IEEE GLOBECOM 2008*, 2008, pp. 1–5.

[7] T. Kim, D. J. Love, and B. Clerckx, "Mimo systems with limited rate differential feedback in slowly varying channels," *IEEE Trans. Commun.*, vol. 59, no. 4, pp. 1175–1189, 2011.

[8] X. Li, N. Park, and Y. Kim, "Differential precoding scheme of lte systems over temporally correlated channels," in *Proc. IEEE Vehicular Technology Conf. (VTC Fall)*, 2011, pp. 1–5.

[9] Y. J. Kim, X. Li, T. Kim, and D. J. Love, "Combination lock-like differential codebook for temporally correlated channels," *Electronics Letters*, vol. 48, no. 1, pp. 45–47, 2012.

[10] B. Clerckx, Y. Zhou, and S. Kim, "Practical codebook design for limited feedback spatial multiplexing," in *Proc. IEEE Int. Conf. Communications ICC '08*, 2008, pp. 3982–3987.

[11] T. Xu and H. Liu, "Index assignment for beamforming with limited-rate imperfect feedback," *IEEE Commun. Lett.*, vol. 11, no. 11, pp. 865–867, 2007.

[12] R. H. Etkin and D. N. C. Tse, "Degrees of freedom in some underspread mimo fading channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1576–1608, 2006.

[13] J. G. Proakis, *Digital Communicaitons*, 4th ed. McGraw Hill, 2000.

 ISBN: 978-1-61208-203-5

High-Throughput Mail Gateways for Mobile E-mail Services based on In-Memory KVS

Masafumi Kinoshita, Gen Tsuchida
Yokohama Research Laboratory
Hitachi, Ltd.
Yokohama-shi, Japan
{masafumi.kinoshita.rt, gen.tsuchida.qc}@hitachi.com

Takafumi Koike
Information & Telecommunication System Company
Hitachi, Ltd.
Kawasaki-shi, Japan
takafumi.koike.kc@hitachi.com

***Abstract*—Mobile network operators providing e-mail services require mail systems to process large volumes of e-mail traffic to and from mobile terminals. Mail gateways, particularly those that accept e-mail messages from external systems and transfer them with store-and-forward communication, require much higher throughput that conventional mail gateways cannot provide. Mail gateways are also required to preserve consistent data and to provide queued services in order. We propose a mail gateway system for mobile e-mail services based on a distributed in-memory key-value-store (KVS) to meet four requirements of high-throughput, high-speed responses, scalability, and availability. We propose KVS to achieve these requirements, which can store messages physically in a queue structure and preserve the consistency of data in respective queues. We present a method of high-throughput access to pipeline messages on an active TCP connection that is linked to a queue on mail gateways and its backup queue in KVS. The mail gateways have a management function for each backup queue in KVS to both preserve consistency and avoid problems in the system. We evaluated the performance of the KVS we propose and a mail gateway corresponding to the KVS. The results proved both the KVS and mail gateway achieved the high throughput that was required.**

Keywords-MTA (mail transfer agent); KVS (key-value-store); in-memory data grid.

I. INTRODUCTION

The growth in the number of users of mobile e-mail services has led to an explosion in the volumes of e-mail traffic encountered by mobile network operators. For example, one mobile network operator has more than 10 million active users and their e-mail system processes more than 10,000 e-mail messages per second.

A large-scale e-mail system is composed of mailbox servers storing the e-mails of their users and mail gateway servers. Mail gateway servers function as mail transfer agents (MTAs), process incoming e-mail messages from external systems, and transfer them to their destinations, such as MTAs on the Internet and mailbox servers. Architectures with mail gateways proposed in [1] are flexible and extendable, and these gateways can stabilize e-mail services by controlling traffic [2]. Mail gateways in mobile e-mail systems also serve to provide billing processes and e-mail security, transcode e-mails, and provide other processes.

Mail gateways generally relay e-mail messages with store-and-forward communication that incoming e-mails are stored in a local queue located within non-volatile storage, which are then forwarded to the destination server. The three main advantages of store-and-forward communication are quick response, control of traffic to avoid burst traffic, and guaranteed e-mail delivery. The main disadvantage of this communication is low throughput because non-volatile storage, such as that in disk and a storage system is accessed, which is a bottleneck in relaying e-mail messages.

Mail gateways in mobile e-mail systems require high throughput, high-speed responses, scalability, and availability. High-throughput and high-speed responses are particularly important for three reasons of: 1) preventing mobile terminals from failing to send e-mails, 2) minimizing connections between mail gateways and mobile terminals to avoid congestion in both e-mail systems and wireless networks, and 3) reducing the number of mail gateway servers. We set the following target values from our experience and expertise as a systems integrator of mobile e-mail service, which has more than 10 million subscribers; high-throughput means a mail gateway should process more than 1,000 e-mails per second, and high-speed response means a mail gateway should respond to a received message within 100 milliseconds.

Well-known MTA software, such as sendmail [3] and postfix [4], fail to meet the high-throughput requirement because their throughput is less than 100 e-mail messages per second [5]. We proposed a method of improving throughput where mail gateways used the method to reduce access disk I/O requests and parallelized these requests while attaching them to storage area network (SAN) storage systems [5]. A mail gateway could process 850 e-mail messages per second and have a response of 80 milliseconds by adopting this method, but this approach failed to meet the requirement for scalability because it was necessary to scale-out too many configurations and too many operations for the storage and servers of mail gateways. Moreover, its performance has recently been insufficient for the requirement for throughput.

Scalability and availability for e-mail systems have also been proposed. Christenson et al. proposed an e-mail system using a network file system (NFS) [7], and Saito et al. [8] and Behren et al. [9] proposed an e-mail system using a distributed storage system with a hash table. Koromilas et al. [11] proposed an e-mail system using Cassandra [10], which is a distributed key-value-store (KVS) software. However,

 ISBN: 978-1-61208-203-5

these proposed systems were mainly designed for the functions of mailboxes, and few considerations were given to mail gateways. These systems, therefore, failed to meet the requirements of high throughput and high-speed responses.

There have been many efforts in the last few years on distributed in-memory KVS or in-memory store technology for high throughput and high-speed responses. A distributed in-memory KVS is composed of multiple nodes, and stores replicated data in the memory of multiple nodes without accessing non-volatile storage. It has been adopted in banking systems [12], stock exchange systems [14], telecommunications [15], and other fields. The systems using it meet the four requirements of high throughput, high-speed responses, scalability, and availability.

Here, we propose a distributed in-memory KVS designed for mail gateways and mail gateways based on it. Four main questions need to be answered in adopting in-memory KVS.

- What is an appropriate architecture for mail gateways for mobile e-mail services to meet all four requirements of high throughput, high-speed responses, scalability, and availability?
- How do they achieve both consistent data and preservation of order for in-order queues for several KVSs? Mail gateways also require these properties for mobile e-mail services.
- How do they efficiently store e-mail messages in KVS to achieve the requirement for high throughput? In other words, the method of storing messages requires high write throughput.
- How do they avoid system problems and recover quickly, without having impact on mobile terminals? In addition, the process for mobile terminals should be executed within several seconds.

The rest of the paper is organized as follows. The background and related work are introduced in Section II. Section III presents the system architecture and design. Section IV describes the implementation and the results obtained from evaluating performance. Section V concludes the paper.

II. BACKGROUND AND RELATED WORK

Fig. 1 outlines an example of the system structure for a mobile e-mail service. One function of mail gateways (GWs) is that they can accept e-mail messages from external systems, such as mobile terminals and MTAs on the Internet, via messaging protocols, such as simple mail transfer protocol (SMTP) and multimedia messaging services (MMS). They can also relay e-mail messages to MTAs or internal mailboxes. Their traffic to MTAs is generally larger than the traffic to mailboxes. Mobile terminals can also access mail gateways to retrieve e-mail messages from the mailboxes, via the Internet message access protocol (IMAP), post office protocol (POP), and MMS.

Mail gateways adopt store-and-forward communication to relay messages via SMTP and MMS. They can also adopt this communication to transfer other message data, such as billing data and notification messages related to e-mails, while processing messages via IMAP and other protocols. Mail gateways with store-and-forward communication store e-mail messages on their disks while waiting for relays. Therefore, they can respond to acceptance of e-mail messages promptly after storing them. This efficiently reduces receipt errors for e-mail messages from mobile terminals, and reduces their impact in both e-mail systems and wireless networks.

Relaying messages may be instantaneous, but this may also be delayed if the destination MTA is unavailable or cannot be reached due to network error. Mail gateways will keep re-trying to make deliveries for a certain period, such as several hours or a few days.

Mail gateways control congestion in MTAs with a destination queue of e-mail messages. Mail gateways also manage queues that have billing data and other messages. Mail gateways are required to manage many queues and have consistent messages in these queues.

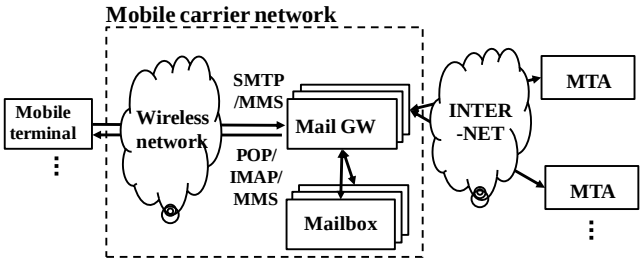


Figure 1. Example of system structure for mobile e-mail service.

We briefly overview related work in what follows. First, we explain studies on conventional mail systems and then we describe some efforts on in-memory KVS.

A. Mail system architecture

Jeun et al. proposed an architecture for a cluster-based e-mail system with an MTA-MDA structure [1] on conventional mail gateways, which equals the mail gateways-mailbox structure discussed in this paper. This architecture was highly scalable, highly available, inexpensive to develop, and had low maintenance costs. Since this system used sendmail [3] and postfix [4] with a network file system (NFS), it did not perform satisfactorily to satisfy the requirement for throughput.

The scalability and availability of e-mail systems have been discussed [8], [9], [11], where scalability has been achieved by using distributed storage systems. The performance of one e-mail system with distributed KVS, i.e., Cassandra [11], is superior to some others, but it is still inadequate to meet the requirement for throughput. Using distributed KVS instead of a conventional storage system provides easy scalability and availability. KVS distributes its data to multiple servers with a consistent hash method [17], [18], which is similar to the methods proposed by Saito et al. [8] and von Behren et al. [9].

Facebook [19] uses distribute KVS, i.e., HBase [20], for its messaging service [21]. They selected it for its scalability, consistency, performance, and other reasons. However, they did not evaluate it.

KVS is a major approach to achieving requirements such as scalability, high performance, and availability.

 ISBN: 978-1-61208-203-5

B. *In-memory store efforts*

Many efforts have been expended on distributed in-memory KVS or in-memory technology for high throughput and high-speed responses. In–memory KVS memcached [16], which is known as high throughput KVS, is used as a cache by many companies, such as Facebook and Twitter [22]. As memcached is a single server and is not replicated in multiple servers, it cannot be a persistent data store. However, distributed in-memory KVS, such as the IBM WebSphere eXtreme Scale [12], can store replicated data in the memory of multiple nodes for persistent storage. They have higher performance than disk-based KVS, such as Cassandra and HBase. Nevertheless, they have two disadvantages compared with disk-based KVS. The first is data are lost if all nodes having replicated data are down at the same time. However, as power supplies are duplicated at the data center and data are backed up to disks periodically, there is little probability that data will be lost. The second disadvantage is that their storage capacity is smaller because memory is more expensive than disks.

As previously mentioned, mail gateways are required to manage queues and keep data consistent. There are a few KVSs storing data in a queue structure, but WebSphere provides a queue service [13]. Its function of queuing was developed as a thin layer on top of a typical KVS structure, which can store one simple set of key values. It stores a meta-data managing queue in KVS, and does not physically store data in a queue structure. However, there are no solutions to resolving the issues with mail gateways described in Section I.

III. PROPOSED ARCHITECTURE AND DESIGN

We first present the mail gateway architecture based on distributed in-memory KVS and the KVS architecture for it in this section. We then propose methods of resolving the issues described in Section I.

A. *Architecture for Mail Gateway System*

The architecture for the mail gateway system is outlined in Fig. 2. There are mail gateway (GW) modules and in-memory KVS modules in each node. These modules are independent of each other, and communicate with other modules in other nodes. The load balancer (L4 switch) dispatches incoming message to mail gateway modules and monitors a TCP port of the mail gateway modules to switch over the path to a non-responding one.

Mail gateway modules receive incoming messages and process them in their memory without accessing their disks. Mail gateway modules back them up (store) in two KVS modules on request. In addition, mail gateway modules also store them in their own queue. Thus, three replicated messages are in memory in respective nodes to avoid them from being lost. Backup in the in-memory only enables high-speed responses.

Mail gateway modules provide consistency and tolerance against partitioning in the consistency, availability and partition-tolerance (CAP) theorem [6]. In addition, availability is important for e-mail services. In this architecture, the availability of a whole system composed of many mail gateway modules is provided by the load balancer switching over a path to mail gateway modules. In addition, there is no single point of failure in the system.

Fig. 2 shows the queue in mail gateway modules are backed up in two KVS modules. KVS modules have a queue structure that physically store messages in queues in their memory. (Details on KVS are described below.) In other words, a queue in mail gateway modules synchronizes two backup queues in a KVS module. Mail gateway modules have many queues of messages to guarantee their order (first in, first out) and they manage the flow to avoid congestion. KVS modules also have the same queues of mail gateway modules for backup. These mail gateway systems attached to KVS have a scalable structure, because they make it easy to add nodes.

Mail gateway modules relate their own queue to backup queues in two KVS modules. They select KVS modules from a group of KVS modules with a set configuration, not using hash distribution on request such as that with Cassandra. Next, they send parameters such as queue length and access accounts, and create backup queues. After that, the mail gateway modules start the service to relay messages with store-and-forward communication. The mail gateway modules store a message in backup queues and its queue. Next, they forward the message to its destination and delete it from the backup queues and its own queue. They occasionally replace messages to back up intermediate states to process messages. These communications (i.e., storing, deleting, and replacing messages) achieve queue data are synchronized between mail gateway modules and KVS modules.

Mail gateway modules monitor backup queues in KVS modules with responses of synchronization and heart-beat health checks. If they detect faults in a backup queue, they isolate it and replicate another backup queue in another KVS. If a mail gateway module's service is down, it reboots, and obtains all data in the backup queue in KVS modules to continue e-mail services. If it cannot reboot, the KVS module moves messages from the backup queue to a program to continue sending them in the same node.

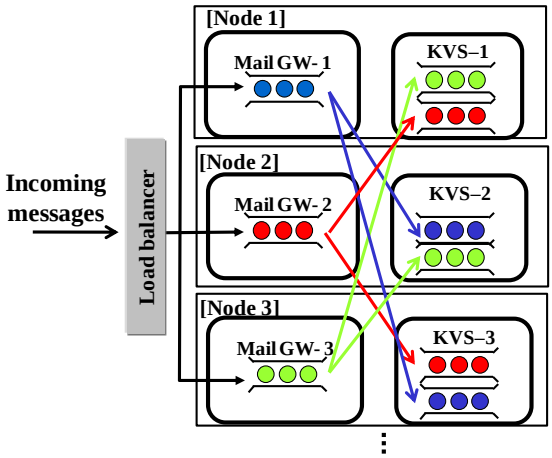


Figure 2. Architecture for mail gateway system.

B. Architecture for KVS with in-order queues structure

As previously mentioned, KVS modules have a queue structure in memory to back up queues in mail gateway modules. Mail gateway modules synchronize their own queue to two backup queues in respective KVS modules. KVS modules support that messages are stored in queues in-order and they retrieve them in-order.

Mail gateway modules have functions of distributing messages to KVS modules having backup queues. KVS modules allow them to access the queue. They can access a message in the structure with a queue's name key and message-id key. This is different from WebSphere [12] in that data in a queue are physically located in respective KVS modules.

This structure where data in queues are physically stored in respective KVS modules has the following advantages. The first advantage is that the number of communication messages is reduced while mail gateway modules and KVS modules are synchronized. That has an impact on the throughput of mail gateways systems. An order index (i.e., metadata for queues) has to be built to preserve the order for in-order queues. The metadata in general KVS, such as WebSphere, are stored as KVS data to preserve queues [12]. Therefore, they have to access a message and several metadata while synchronizing one message. The proposed KVS modules, on the other hand, physically have the metadata as a table in their own memory, and they only have to access one message while synchronizing themselves.

The second advantage is that queue data can be quickly recovered if there is trouble, such as network error or nodes are down. For example, if a mail gateway module reboots, it retrieves all data in queues from a KVS module synchronized with it. It can immediately obtain a block including many messages in queues. If a typical KVS were used instead of the proposed KVS modules, mail gateways could only obtain one message immediately and they would have to access metadata in the queue every time they obtained it. In addition, since these messages are distributed to KVS (respective nodes) with the hash table, mail gateways access many nodes with respect to each message and metadata. Thus, the queue structure can significantly reduce recovery time.

The third advantage is the availability of mail gateway systems. As previously mentioned, their availability is provided by the load balancer switching over a path to mail gateway modules. Mail gateway modules without synchronized KVS modules close their own TCP ports and suspend their own services. If a KVS module has some fault, the load balancer can definitely isolate mail gateway modules by closing their ports. If typical KVS were used instead of the proposed KVS modules, all mail gateway modules could access these KVS modules while faulty KVS modules would not be isolated, and the impact of these faults would spread to whole systems.

C. High-throughput method to access KVS

To preserve order in in-order queues and achieve the requirement for high throughput, mail gateway systems adopt the following method of communication. Fig. 3 outlines the flow for the method of communication between a mail gateway module and a KVS module. A mail gateway communicates with a KVS module with persistent transmission control protocol (TCP) connections. One queue of a mail gateway module and one queue of a KVS module are linked with two connections; these are via a respective path in different network segments. One connection is active for the communication message, and the other is on standby when network error occurs in the active path. Since one TCP connection is used to achieve communication in order, the order in the queue is preserved.

First, a mail gateway module connects to a KVS module with an authentication message, and it sends a request creating a queue whose parameters, e.g., queue length and access accounts, are requested by it. After a KVS module accepts the request, it creates a queue in its memory and this queue is related to the connection in its management table of queues. This sequence means the mail gateways have obtained ownership of the queue.

The mail gateway module sends pipelined-requests on an active connection to the KVS module. The KVS module immediately receives pipelined requests and processes them. Finally, the KVS module immediately sends pipelined responses on the connection. The KVS module obtains an internal queue lock while processing messages from the connection. The requests and the responses include a "transaction number", i.e., the number it has already processed, and a "sequence number", i.e., the number it requires for matching requests and responses. The number of transactions is checked to preserve the order in queues.

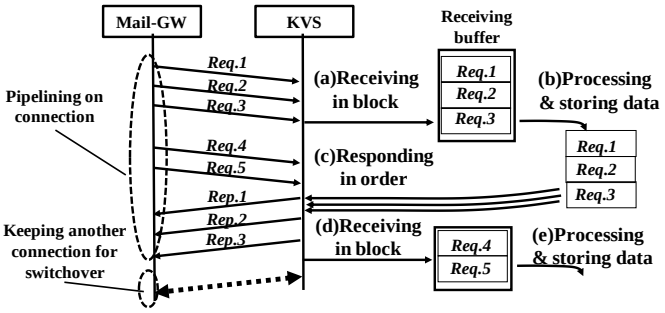


Figure 3. Method of comunication between mail gateways and KVS

Thus, this method of communication can provide high-throughput and consistent storage in queues. Typical KVS, such as memcached, Cassandra, and Dynamo are generally used in systems that have many low-performance clients that access KVS with many connections. However, the mail gateway modules in the proposed system are high performance and there are an equal number of mail gateway modules and KVS modules. There are two ways of efficiently communicating data via networks; the first is by communicating with many connections, and the second is by multi-communicating with several connections. The former in this mail gateway system can decrease performance because KVS modules control (obtain and release) internal locks by processing one message from each connection. The latter reduces the number of locks that are controlled.

 ISBN: 978-1-61208-203-5

However, the latter is susceptible to the control of TCP flows because there are large message traffic flows on the connections. There is the method of avoiding this impact, i.e., a mail gateway module can use a sufficient number of queues to adjust message traffic on the connection. Therefore, we selected the latter, which we evaluated and explain below.

The above issue is similar to those with the method of communication between parallel and serial communications. In other words, our proposed method of communication between mail gateway modules and KVS modules is similar to serial communication on TCP connections.

D. Mail Gateway's Management of KVS

Mail gateway systems require consistency for e-mail services. Mail gateway modules have a management table for each backup queue in KVS modules to preserve consistency (summarized in Table. I). They manage two states of one queue, i.e., a "network state" and a "synchronized state". The network state indicates the state of a network path. Network state (A) means the state of the active connection, and (S) means the state of the standby one. The network state is updated in these two cases, where mail gateways are connecting and network error is caused during synchronization or heart-beat health checks. Mail gateway modules switch over to standby connection in seconds when active connections are unavailable.

The synchronization state indicates whether the queue is synchronized or not. Mail gateway modules check the number of transactions to monitor the synchronization state, (which is the number KVS has already processed to preserve the order of queues), in responses from KVS. A mail gateway module uses this table in synchronization, heart-beat health checks, and other processes. If the synchronization state is no good (NG), a mail gateway deletes all data in this backup queue in KVS and re-synchronizes the data in the entire queue.

When mail gateway modules store a message in a queue, they send a message to the KVS modules indicated in this table. They succeed in storing the message after receiving all responses from KVS modules. If they cannot receive all responses before timeout, which is seconds, they retry to send the message on a standby connection and change the network state. Finally, when they fail to store it in KVS, they isolate this backup queue, and create a backup queue in another KVS by using any conditions.

TABLE I. MANAGEMENT TABLE FOR BACKUP QUEUES

Queue Name (dst.domain)	KVS	Network State (A)	Network State (S)	Sync. State	Number of Transactions
hitachi.com	NodeB	Run	Ready	Sync	18000
	NodeC	NG	NG	NG	-
xxxxxx.com	NodeD	NG	Run	Sync	3000
	NodeE	NG	Run	Sync	3000

IV. IMPLEMENTATION AND EVALUATION

We first present the implementation of the proposed system in this section and the methodology we used to evaluate it. We evaluated the throughput of one KVS module, the throughput of mail gateway modules corresponding to two KVS modules, and the process time for recovery.

A. Implementation and methodology for evalution

Mail gateway modules and KVS modules were implemented with event driven architecture [23] developed in the C language. We designed their performance, especially throughput that could be scaled up with increasing CPU frequency.

Mail gateway modules support SMTP, IMAP, and other protocols. Mail gateway modules can have 1,000 queues at maximum. Although the length of a queue can be configured for more than a million messages, it is limited by the memory size of each module. These queues are used for storing e-mail messages, billing their data, and storing and forwarding their notification messages and other messages. The queues for e-mail messages via SMTP services can also provide several functions; a control function for the e-mail traffic of each destination MTA or mailbox server, a control function for rates, which shows how many messages to send per second, a function for regulating the receipt and sending of e-mail messages, and a function for timeout to transfer them.

First, we evaluated the transaction throughput for a single KVS module. Second, we evaluated throughput via the SMTP of mail gateway modules corresponding to KVS modules. Next, we evaluated how quickly the system recovered from server incidents.

B. Throughput of proposed KVS

There were two nodes that had dual processor dual cores and 4 GB of RAM, and they were connected to two Gigabit Ethernet networks. One node includes a KVS module, and the other included a test program. The test program generated the workload to the KVS modules, received responses, and evaluated throughput, i.e., the number of transactions per second. One transaction was a set where a message was stored and deleted because the transaction of relaying an e-mail message with storage and forwarding includes a set.

We evaluated the KVS module for different-sized messages from 0.4 to 20 KB. The reason we evaluated these sizes for messages is that they represent the greatest volume of e-mail traffic in mobile e-mail services from our experience. The KVS module had one queue and the test program sent messages to the queue with one active connection.

We also evaluated memcached [16] to compare it with the evaluation of the proposed KVS module. Memcached has a very simple KVS that does not have lock control, a queue structure, or other functions; therefore, its throughput is known to be high.

Fig. 4 plots the transaction throughput for a single KVS for different sized messages. The throughput for the proposed KVS modules for 0.4 KB is 200,000 transactions per second, which is twice the throughput of memcached. The throughput for the proposed KVS for 1 KB is 100,000 transactions per second, which is 1.4 times the throughput of

 ISBN: 978-1-61208-203-5

memcached. The throughput for the proposed KVS for more than 2 KB is limited by the 1-Gbps network, just like it is for memcached. Thus, the experimental results proved the proposed KVS modules met the requirement for high throughput to store messages. In addition, it proved the method of communicating messages with pipelining on one active connection was efficient.

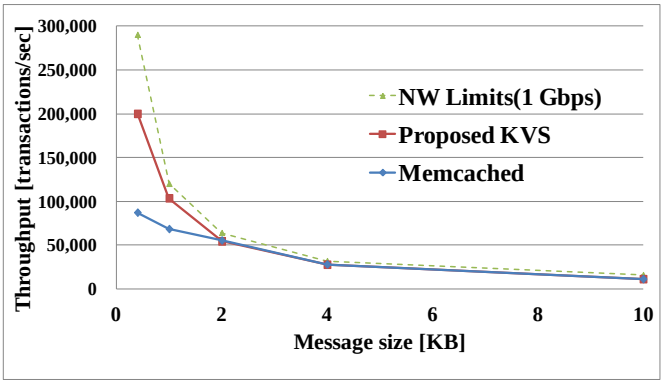


Figure 4. Transaction throughput of KVS for different sized messages

C. *Throughput of mail gateways attached to KVS*

There were two nodes that had six processor dual cores and 32 GB of RAM, and they were connected to two Gigabit Ethernet networks. One node included a mail gateway module, and the other included two KVS modules and two kinds of test programs. All programs are connected via Gigabit Ethernet network. The first was a test client program that generated the workload of e-mail messages to the mail gateway modules via SMTP, received responses, and evaluated throughput, i.e., the number of messages per second. The second was a test server program that received e-mail messages from the mail gateway module. The mail gateway module received an e-mail message from the test client program and stored it in backup queues in two KVS modules. After that, the mails gateways sent a response to the message to the test client program and concurrently transferred it to the test server program.

We evaluated a mail gateway with a workload composed of 70 percent 1KB-messages and 30 percent 10KB-messages, to simulate realistic message traffic for mobile e-mail services.

We compared our evaluations of the proposed method using KVS modules with a conventional method where mail gateways used a redundant array of independent disks (RAID) storage with a method of streamlining disk I/O requests [5]. Accessing disks to store messages is a bottleneck for throughput when the conventional method is adopted. A past experiment by Kinoshita et al. [5] adopted the conventional method to compare its performance with well-known MTA software; the throughput for sendmail [3] was less than 20 messages per second, and the throughput for postfix [4] was 80 messages per second.

Fig. 5 compares the throughput and average response times of the proposed and conventional methods. The throughput for the proposed mail gateway module is 3,600 e-mail messages per second, which is 4.4 times the throughput of the conventional method. The throughput is less than that of a single KVS, i.e., 11,000 transactions per second. This means accessing KVS modules no longer creates bottlenecks in throughput. The average response time for the proposed mail gateway module is 14 milliseconds, which is an fifth of the response with the conventional method. Thus, the experimental results prove that the proposed mail gateway module meets both the requirements of high throughput and high-speed response described in Section I.

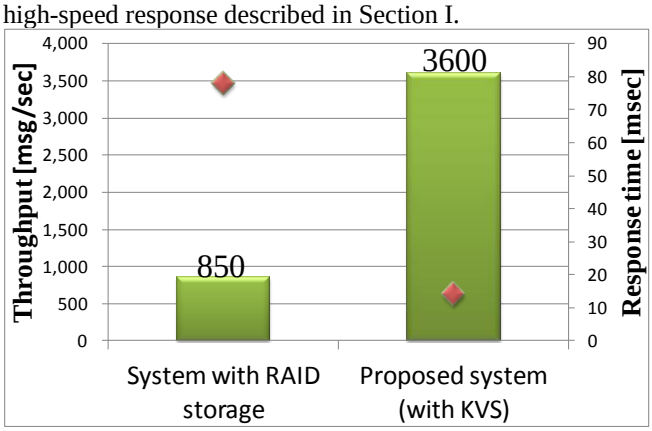


Figure 5. Throughput and average response time.

D. *Time for Recovery*

As previously mentioned, one advantage of the proposed system is its quick recovery of queue data. A mail gateway module or a KVS module can reduce the time for recovery to immediately obtain a block that includes many backup messages for queues. We evaluated the time for recovery for different numbers of messages in the block.

There were two nodes that had six processor dual cores and 32 GB of RAM, and they were connected to two Gigabit Ethernet networks. One node included a mail gateway module, the other included KVS modules and a stored e-mail program. First, the program stored messages in the mail gateway module. The mail gateway module stored a number of messages in the backup queues in two KVS modules and its own local queue. We then shut down the mail gateway module to simulate the server down, and rebooted it. After that, the mail gateway module obtained all its messages from a backup queue in a KVS module. This process meant recovery from trouble with mail gateway modules.

We evaluated the time it would take for this recovery process where the mail gateway module obtained all messages from a backup queue in a KVS module. The message size was 1 KB and the queue had 500,000 messages.

The experimental results are given in Fig. 6, which prove that the more messages there are in a block, the shorter the recovery time is. The recovery time for 1,000 messages was 5.9 seconds, which is a fourteenth of the recovery time for one message. The case of one message in a block was the same as the method of recovery for a typical KVS, such as memcached or WebSphere (with a thin layer of queue service implemented) [12]. In addition, if a typical KVS is used instead of the proposed KVS modules, there is more than double the access to the metadata of queues in KVS.

 ISBN: 978-1-61208-203-5

Therefore, the recovery time for using typical KVS is more than 160 seconds; which cannot be adopted for mobile e-mail services. Thus, the experimental results proved the proposed system met the requirement for rapid recovery described in Section I.

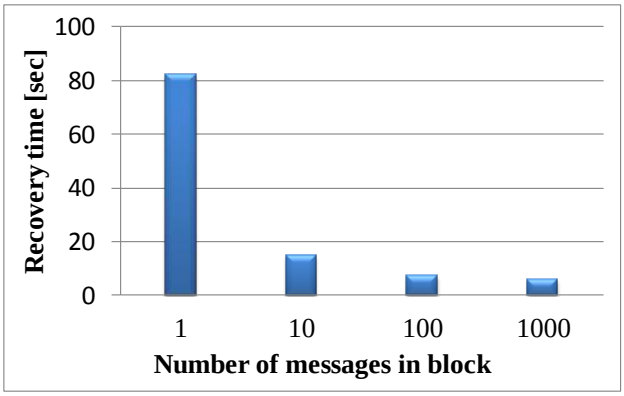


Figure 6. Recovery time for different numbers of messages in block

E. Potential for Lost Messages

This system stores messages only in the memory of multiple nodes. If all nodes having replicated data are down at the same time, this system loses messages. Therefore, it has to be located in a data center that has a stable power supply. The power supplies of all nodes are generally duplicated in the data centers of mobile network operators. In addition, mail gateway modules have to store messages for several hours, or days at maximum. Thus, there is a very low probability of messages being lost in this system.

V. CONCLUSION

We presented a mail gateway system based on distributed in-memory KVS to satisfy the four requirements of high throughput, high-speed responses, scalability, and availability for mobile e-mail services.

We proposed KVS, which can store messages physically in a queue structure and provide consistent data in respective queues. To preserve order in in-order queues and achieve the requirement for high throughput of backup to KVS, we proposed a method of pipelining messages on an active TCP connection that linked a queue in mail gateway modules and its backup queue in KVS modules. In addition, mail gateway modules switched over to standby connection in different network segments in seconds when active connection was unavailable. The mail gateway modules had a management table for each backup queue in KVS to avoid system problems. The system recovered almost immediately within seconds to obtain a block including many backup messages for queues.

We evaluated the performance of one KVS module, the performance of a mail gateway module that corresponded to two KVS modules, and the process time for recovery. The results proved both the KVS modules and the proposed system met the requirement for high throughput. The throughput for the proposed KVS module was 200,000 transactions per second with 0.4-KB messages, which is superior to memcached. The throughput for a proposed mail gateway module was 3,600 e-mail messages per second, which is 4.4 times the performance of the conventional method. Since the proposed KVS was adopted for mail gateways instead of conventional storage, the process of persistently storing messages was no longer a bottleneck to throughput.

REFERENCES

[1] W.-C. Jeun, Y.-S. Kee1, J.-S. Kim, and S. Ha, " High Performance and Low Cost Cluster-Based E-mail System", Parallel Computing Technologies, pp. 482–496, 2003.

[2] M. Grubb., "How to Get There From Here: Scaling the Enterprise - Wide Mail Infrastructure", In the Proceedings of the Tenth USENIX Systems Administration Conference (LISA '96), Chicago, pp. 131–138, 1996.

[3] sendmail: http://www.sendmail.com/sm/open_source/ April.6.2012

[4] postfix: http://www.postfix.org/ April.6.2012

[5] M. Kinoshita, M. Nakahara, and T. Sagara, "An Implementation and Evaluation of Multiprotocol Message Gateway", The 71th National Convention of IPSJ , March 2009.

[6] E. A. Brewer, "Towards robust distributed systems", In Proceedings of the 19th Annual ACM Symposium on Principles of Distributed Computing, p. 7, 2000.

[7] N. Christenson, T. Bosserman, and D. Beckemeyer, EarthLink Network, Inc., "Highly Scalable Electronic Mail Service Using Open Systems", Proceedings of the USENIX Symposium on Internet Technologies and Systems, Monterey, California, December 1997.

[8] Y. Saito, B. N. Bershad, and H. M. Levy, "Manageability, availability and performance in Porcupine: A highly scalable, cluster-based mail service", 17th ACM Symposium on Operating System Review, 34 (5) pp. 1–15, 1999.

[9] J. R. von Behren, S. Czerwinski, A. D. Joseph, E. A. Brewer, and J. Kubiatowicz, "NinjaMail: the Design of a High-Performance Clustered, Distributed E-mail System", In Proceeding of International Workshops on Parallel. Processing 2000, pp. 151–158, 2000.

[10] apach cassandra: http://cassandra.apache.org/ April.6.2012

[11] L. Koromilas and K. Magoutis, "CassMail: A Scalable, Highly-Available, and Rapidly-Prototyped E-Mail Service", Lecture Notes in Computer Science, Volume 6723/2011, pp. 278–291, 2011.

[12] IBM WebSphere eXtreme Scale: http://www-01.ibm.com/software/webservers/appserv/extremescale/ April.6.2012

[13] Y. Wang, H. Chen, B. Wang, J. M. Xu, and H. Lei, "Scalable Queuing Service Based on an In-Memory Data Grid", IEEE 7th International Conference on e-Business Engineering (ICEBE 2010), pp. 236–243, November 2010.

[14] Y. Hashidume, K. Takasaki, T. Yamazaki, and S. Yamamoto, "Ultra-high-speed In-memory Data Management Software Achieving High-speed Response and High Throughput", Fujitsu Scientific & Technical Journal, Vol. 62, pp. 57–64, January 2011.

[15] S. Kondoh, Y. Miyagi, M. Kaneko, T. Fukumoto, and K. Ueda, "A Study of Data Arrangement for Various Retrieval and Effective Redundancy", IECE Technical Report, NS2011-63, pp. 11–16, September 2011.

[16] memcached: http://memcached.org/ April.6.2012

 ISBN: 978-1-61208-203-5

[17] G. DeCandia et al. ,"Dynamo: Amazon's Highly Available Key-value Store", Proceedings of 21st ACM SIGOPS Symposium on Operating Systems Principles (SOSP'07), pp. 205–220, October 2007.

[18] A. Lakshman and P. Malik, "Cassandra -A Decentralized Structured Storage System", Cornell, 2009.

[19] Facebook: http://www.facebook.com/ April.6.2012

[20] HBase: http://hbase.apache.org/ April.6.2012

[21] D. Borthakur et al., "Apache hadoop goes realtime at Facebook", Proceedings of the 2011 International Conference on Management of Data, SIGMOD '11, pp. 1071–1080, 2011.

[22] Twitter: http://www.twitter.com/ April.6.2012

[23] M. Welsh, D. Culler, and E. Brewer, "SEDA: An architecture for well-conditioned, scalable Internet services.", In Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP '01), ACM Press, pp. 230–243, October 2001.

 ISBN: 978-1-61208-203-5

Transmission of JPEG2000 Images in an Uplink Cellular Network with UPA and SCFDE: A System Description

Moein Shayegannia
Dept. of Engineering Science
Simon Fraser University
Burnaby, Canada
moeins@sfu.ca

Sami Muhaidat
Khalifa University of Science Technology & Research
Abu Dhabi, UAE
muhaidat@ieee.org

Atousa HajShirMohammadi
Dept. of Engineering Science
Simon Fraser University
Burnaby, Canada
atousah@sfu.ca

***Abstract*—In this paper, we describe a work in progress for transmission of JPEG2000 images using an unequal power allocation algorithm and single carrier frequency domain equalization technique over a block fading frequency selective channel. The optimization algorithm exploits the hierarchical structure of the JPEG2000 images and uses a distortion model along with the channel state information for allocating optimal values of power on each coding pass to minimize the end-to-end distortion. The single carrier frequency domain equalization technique combats the negative impact of inter symbol interference and inter carrier interference in a frequency selective channel. Two antennas at the transmission terminal are utilized to evaluate the performance of our system in a multi input single output scenario using space time block coding scheme. The simulation results present that integration of the unequal power allocation technique with orthogonal frequency division multiplexing yields a higher quality than when single carrier frequency domain equalization is used instead. In addition, the transmit diversity technique enhances the peak signal to noise ratio of the received image for about 2.5 dB. These results prove that our power allocation algorithm is more compatible with orthogonal frequency division multiplexing than single carrier frequency domain equalization, and it yields spacial diversity in a multi transmitter antenna system.**

Keywords- JPEG2000; unequal power allocation; single carrier frequency domain equalization; space time block coding; wireless image transmission.

I. INTRODUCTION

Transfer of multimedia streams with high reliability of the received signal, and high data rate over wireless channels is becoming more attractive in the current mobile era. The increasing demand of access to multimedia data and the hostile behavior of wireless medium impose challenges in maintaining the total available bandwidth, the total transmission power, and quality of the transmitted stream. To address these challenges, we need to alleviate the sources of disturbance in wireless channels, such as fading, interference, shadowing, path loss, and multipath, by providing an effective data protection technique.

The inherit redundancy contained within multimedia signals and bandwidth limitation require compression of image and video streams before transmission. JPEG2000 is one of the recent and advanced source coding techniques for image coding. This standard provides some degrees of protection against errors in noisy channels with the help of an error-resilient feature. The error resilient tool is able to detect errors but it can not correct any of them in the code-stream. Once an error is detected, the rest of data is discarded and resynchronization mechanism of the received data is performed [1]. Therefore, the error resilience tool of JPEG2000, by its own, is insufficient for alleviating sever channel impairments. A higher degree of protection is achievable with the aim of the hierarchical structure of JPEG2000 coded bitstream in which some bits hold more important information compared to others. This scalable coded bitstream enables us to provide higher protection over the more important bits. Different techniques such as Unequal Error Protection (UEP) and Unequal Power Allocation (UPA) are introduced in the literature to enhance the transmission of scalable coded bitstreams. The UEP methods apply Forward Error Correction (FEC) coding with different coding rates to different portions of the bitstream, based on the importance of each portion. The UPA techniques distribute the total available power for transmission of an image unequally over the bitstream in such a way that more power is allocated to the more important bits.

Transmission of JPEG2000 images using UEP techniques over slow fading non-frequency selective channels has been widely investigated in [2]-[5]. In [2], UEP is achieved in JPEG2000 code-streams by using Reed Solomon (RS) block codes. In [3], Banister *et al.* make use of Viterbi algorithm to jointly optimize source rate and channel rate for the purpose of UEP. In [4], the authors employ product coded streams which consist of Turbo-codes and RS codes to obtain UEP. In [5], the authors obtain UEP using RS channel coding for the header and convolutional coding for the body of the image bitstream. Protection of JPEG2000 images over slow fading non-frequency selective channels using UPA schemes has been reported in [6] and [7]. Atzori employs an optimized UPA scheme in [6] based on increasing JPEG2000 image quality as well as RS channel coding to protect coded bitstream. In [7], we proposed an optimized UPA scheme based on minimizing the total end to end distortion of JPEG2000 images, which proved its effectiveness in improvement of the Peak Signal to Noise ratio (PSNR) performance and at a lower complexity in comparison with the existing UEP techniques.

Evaluation of UEP for JPEG2000 image transmission over frequency selective channels is investigated in [8]-[11]. Houas *et al.* prove the efficiency of their UEP technique with rate compatible punctured convolutional codes in an OFDM system [8]. In [9], the layer structure of the JPEG2000 is exploited by protecting data in top layer with an FEC code, and the performance is analyzed in an OFDM transmission system. In [10], authors achieve UEP with the means of an optimal joint source-

 ISBN: 978-1-61208-203-5

channel coding and consider progressive image transmission over differentially space-time coded OFDM system. Sethakaset *et al.* propose an UEP scheme in the spatial domain through enhanced beamforming algorithms over closed loop multiple input multiple output OFDM system [11]. Despite the extensive reports on the performance evaluation of UEP techniques for image transmission over frequency selective channels, effective performance of an optimized UPA scheme for transmission of JPEG2000 images in frequency selective channels has not been analyzed. To address this issue, we prove the efficiency of our proposed optimized UPA scheme for frequency selective channels in [12] by combining the UPA algorithm with the OFDM technique.

All the literature reports on multimedia data transmission over frequency selective channels, with UEP or UPA, make use of OFDM to combat the negative impact of Inter symbol Interference (ISI) and Inter Carrier Interference (ICI). The multicarrier implementation of OFDM technique leads to several drawbacks including large Peak-to-Average Power Ratio (PAPR) and high sensitivity to carrier frequency offsets [13]. To tackle these issues, OFDM is implemented in the base station of a cellular network for downlink data transmissions, and Single Carrier Frequency Domain Equalization (SCFDE) is adapted at the end user station for uplink data transmissions. SCFDE is comparable with OFDM in terms of complexity, while it avoids the drawbacks associated with OFDM [13]. In this paper, we integrate our proposed UPA scheme with SCFDE to provide protection for uplink transmission of JPEG2000 coded bitstreams in block fading frequency selective channels. Moreover, we obtain full spatial diversity in our transmission by using Alamouti's Space Time Block Coding (STBC) scheme [14]. Continuation of this work includes further investigation on the impact of SCFDE and STBC on the bit budget of the JPEG2000 coded bitstream for the possible amount of conservation on the bandwidth.

The rest of this paper is organized as follows: Section II provides a brief overview of JPEG2000 image coding, Section III presents the system model, the simulation results are presented in Section IV, and Section V concludes the paper.

Notations: $[\cdot]^H$ and $|\cdot|$ denote the Hermitian transpose and the absolute value Operations, respectively.$[\cdot]_{ik}$ refers to the $(i,k)^{th}$ entry of a matrix. $[\cdot]_i$ means the i^{th} entry of a vector. $\mathbf{Q}$ represents an $N \times N$ FFT matrix whose $(i,k)^{th}$ element is given by $1/\sqrt{N}exp\ (-j2\pi ik/N)$ where $0 \leq i,k \leq N-1$. Bold uppercase letters denote matrices, bold lowercase letters represent vectors and lowercase letters denote scalar variables.

II. Review of JPEG2000 Image Coder

In the JPEG2000 image coder, the first operation is to (optionally) partition a source image into a number of rectangular non-overlapping blocks called tiles. Then Discrete Wavelet Transform (DWT) is applied to the tile which transforms the samples into spacial frequency subbands at different levels of resolution. The first level of decomposition consists of four subbands LL1, LH1, HL1, HH1 [15]. The LL1 subband is the lowest resolution of the tile and is a down-sampled low-resolution representation of the original tile-component. The LL1 subband can be further decomposed by applying DWT. This process can be repeated to obtain different resolution levels. Then, each resolution of each tile component is further partitioned into precincts. Within every subband, each precinct contributes one packet to the code-stream

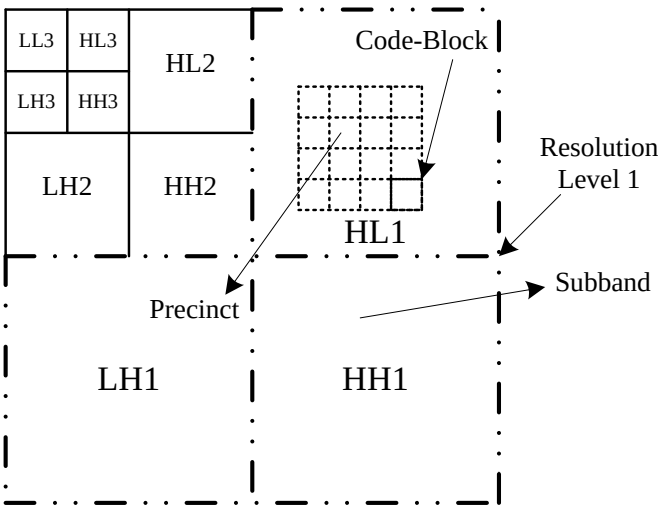


Fig. 1: Components of a JPEG2000 transformed image

of the image. Precincts are not a partition of image data and do not impact sample data transformation or coding. Precincts are used to reconstruct the resolution [2].

Each subband is partitioned into small blocks, called code-blocks, where quantization and bit-plane coding are initiated. The packets furnished by the precincts identify their header and body from the contributions of the code-blocks belonging to the relevant precinct. Starting from the Most Significant Bit (MSB), the coder scans through the bit-planes of each coding pass. Each of the coding passes collects the relevant information about the bit-plane data. Based on the significance of a particular bit location and its neighboring, location of each bit in the coding passes is decided. The encoder uses this information to generate a compressed bitstream or a code-stream [1]. Fig.1 illustrates a 3 layer decomposition of a source image using DWT and its partitioning into four resolution levels, subbands, precincts, and code-blocks.

To increase the robustness of the JPEG2000 bitstream against error propagation along the code-stream, error resilient feature is introduced in the standard. Small size code-blocks are independently coded and included with resynchronization markers. As a result, errors do not propagate beyond the code-block whose bit-stream is corrupted, and the markers keep synchronization between the encoder and decoder in case of occurrence of bit errors. JPEG2000 standard also provides a mechanism to combine all the packet headers within the main header. This adds an advantage to the decoding process of the received data stream, if the main header can be transmitted in an error-free medium. It is necessary to correctly decode the header of a packet in order to extract the code-block contributions to the body of the packet [1], [15].

III. System Model

The overall system block diagram that we are implementing in this study is shown in Fig. 2. We have presented detailed explanations on the functionality of the JPEG2000 encoder, Structure Information Retrieval, UPA optimization, and the Power Adjustment units in [7]. For the sake of completeness, we present a short overview of these units here. The JPEG2000 encoder transforms the format of an input image into JPEG2000 format and generates a scalable coded bitstream. The Structural

 ISBN: 978-1-61208-203-5

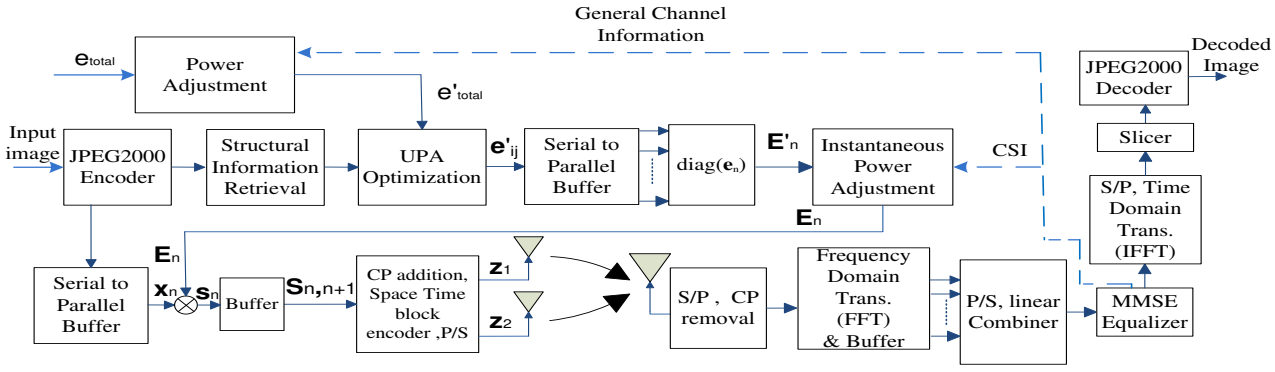


Fig. 2: System block diagram

Information Retrieval unit recovers the required information, such as the number of code-blocks and the number of coding passes within each code-block, from the source coder and passes them to the UPA optimization unit. In the UPA optimization unit runs Simulated Annealing optimization algorithm on the coded bitstream of the JPEG2000 image. This algorithm aims at minimizing the total end to end distortion of the received image by allocating optimal amount of power to each coding pass. During the process of power allocation, the UPA optimization unit obtains the total available power to distribute among the coding passes ($e^{'}_{total}$) from the power adjustment unit. The UPA optimization unit assumes an Additive White Gaussian Noise (AWGN) channel for the transmission medium. The fading impact of the channel is taken into consideration in the Instantaneous Power Adjustment unit by altering the amount of power allocated to each coding pass by the UPA algorithm. In part C, we will describe the functionality of the power adjustment unit. The bitstream and the power assigned to each bit are categorized into N_B blocks. Before the transmission instant, every block of data is appended with cyclic prefix, and two consecutive blocks , n^{th} and $(n+1)^{th}$ where $(n = 1, 2, ..., N_B)$, are passed to the Space Time Block Encoder. The received terminal includes the model of a typical SCFDE receiver.

A. Channel Model

The wireless communication channel is considered to be block fading frequency selective, where two antennas are used at the transmitter side, and one antenna at the receive terminal. First, we will elaborate the transmission model in the case of one transmit antenna, and then we will expand it to two transmit antennas in part D. The channel impulse response for the n^{th} transmission block is given by $\mathbf{h}_n = [h_0 \ h_1 \ ... \ h_{m-1}]^T$ where m is the channel memory length and $n = 1, 2, ..., N_B$. We append the beginning of each transmitting block of data with a size of $N \times 1$ with the last m samples of the same block to eliminate the impact of Inter Block Interference (IBI). At the receiver side, the first m samples of every block is discarded and only N samples are processed. We can account for the addition and removal of the cyclic prefixes by forming the channel into a circulant matrix for every n^{th} transmission block. Thus, assuming a single transmit antenna, the received signal is given by

$$\mathbf{r}_n = \mathbf{H}_n \sqrt{\mathbf{E}_n} \mathbf{x}_n + \mathbf{v}_n \quad n = 1, 2, ..., N_B \tag{1}$$

where $\mathbf{r}_n$ is the n^{th} received block of data, $\mathbf{H}_n$ is an $N \times N$ circulant matrix for the n^{th} transmission block with entries $[\mathbf{H}_n]_{ik} = [\mathbf{h}_n]_{(i-k) \ mod \ N}$ and $[\mathbf{h}_n]_i = 0$ for $i > m-1$. $\mathbf{v}_n$ is an $N \times 1$ additive white Gaussian noise vector with mean of zero and variance of 1/2 per dimension. Since $\mathbf{H}$ is a circulant matrix, it can be eigen-decomposed to form

$$\mathbf{H}_n = \mathbf{Q}^H \mathbf{\Lambda}_n \mathbf{Q} \tag{2}$$

$\mathbf{\Lambda}_n$ is a diagonal matrix of size $N \times N$ for the n^{th} transmission block, in which the diagonal elements are [16]:

$$[\mathbf{\Lambda}_n]_{ii} = \sqrt{N} \mathbf{q}_i^H \left(\mathbf{h}_n \overbrace{0 \ 0 \ ... \ 0}^{N-m} \right)^T \quad i = 1, 2, ..., N \tag{3}$$

where $\mathbf{q}_i$ is the i^{th} column of the matrix $\mathbf{Q}$. The channel impulse response varies for every transmission block of data, and so does the diagonal elements of $\mathbf{\Lambda}$. These elements are in fact the eigenvalues of the channel matrix ($\mathbf{H}$).

B. Single Carrier Frequency Domain Equalization Model

In [17], a thorough overview of SCFDE with STBC scheme is presented, and we use this scheme to form our receiver model. At the receiver side, initially a serial to parallel conversion is performed and then the redundant cyclic prefix data are discarded. As known from its name, the equalization of SCFDE is carried on in the frequency domain. Thus, we transform the received time domain block $\mathbf{r}_n$ into frequency domain by applying the FFT

$$\mathbf{Q}\mathbf{r}_n = \mathbf{Q}\mathbf{H}_n \sqrt{\mathbf{E}_n} \mathbf{x}_n + \mathbf{Q}\mathbf{v}_n = \mathbf{\Lambda}_n \sqrt{\mathbf{E}_n} \mathbf{x}_n + \mathbf{Q}\mathbf{v}_n \tag{4}$$

The next step is to equalize the signal using Minimum Mean Square Error (MMSE) estimator. Our MMSE-SCFDE for the n^{th} received block is given by a diagonal matrix of size $N \times N$ with the following elements [17]:

$$[\mathbf{W}_n]_{ii} = \frac{[\mathbf{\Lambda}_n^H]_{ii}}{|[\mathbf{\Lambda}_n]_{ii}|^2 + \frac{1}{SNR}} \tag{5}$$

where SNR is the Signal to Noise Ratio. The output of the MMSE equalizer will be $\mathbf{y}_n = \mathbf{W}_n \mathbf{r}_n$ and can be transferred to time domain by applying the IFFT to recover the original data given by $\hat{\mathbf{x}}_n = \mathbf{Q}^H \mathbf{y}_n$. Then we can apply a hard decision slicer on the recovered data, and pass it to the JPEG2000 decoder to generate the received image.

 ISBN: 978-1-61208-203-5

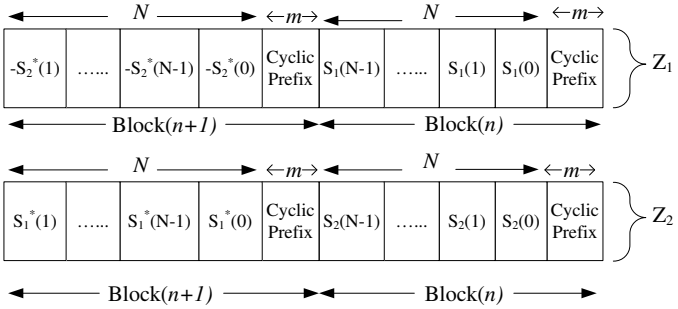


Fig. 3: Transmission block format for SCFDE-STBC [17]

C. Instantaneous Power Adjustment Unit

Our proposed UPA algorithm in [7] assumes an AWGN channel at the time of optimal power allocation to the coding passes. The Instantaneous Power Adjustment unit compensates for the effect of fading by using the instantaneous values of the channel impulse response

$$\mathbf{E}_n = \frac{\mathbf{E}_n^{'}}{\alpha + \left|\sum_{k=0}^{m-1} h_k\right|^2} \tag{6}$$

where h_k is the k^{th} tap channel impulse response for the n^{th} transmission block and m is the channel memory length or the total number of taps. $\mathbf{E}_n^{'}$ is the assigned power to the bits within the n^{th} transmission block and $\mathbf{E}_n$ represents the corresponding power of the bits after the effect of channel fading is taken into consideration. α is small constant value which prevents new zero values in the denominator when the channel is sever and the fading factor is small. In addition, this constant value avoids very large adjusted power to keep the power amplifiers perform in their linear region. For this part, we assume that the Channel State Information (CSI) is available at the receiver side and can be communicated to the transmitter. There are plenty of papers in the literature that report on the estimation of CSI using pilot assisted techniques or blind estimation methods [18], [19].

D. Space Time Block Coding for Transmit Diversity Scheme

In our system design, we employ two antennas at the transmitter side to benefit from spacial diversity. This increases the reliability of the wireless link and improves the quality of the received image. We can also conserve energy by maintain similar received image quality as the single transmit antenna case, however at a lower transmit power. To explain the functionality of the space time block encoder unit and the linear combiner unit in Fig. 2, we follow the proposed transmit diversity methodology in [17]. The two antennas, deployed at the transmitter terminal in Fig. 2, send two transmitting blocks of data ($\mathbf{z}_1$ and $\mathbf{z}_2$) to the channel. Figure 3 presents the structure of these data blocks. In this structure, cyclic prefixes are appended to each block and then discarded at the receiver side in order to eliminate IBI and shape all the channel matrices circulant. An important assumption that we consider in this part for the transmission of the STBC structures is that the channel impulse response remains constant over two consecutive blocks of data. The available power for transmission at any of the two antennas is half of its value in the single transmit case. This helps to keep the total transmit power constant. The linear combiner methodology is detailed in [17] and we will follow the same principle.

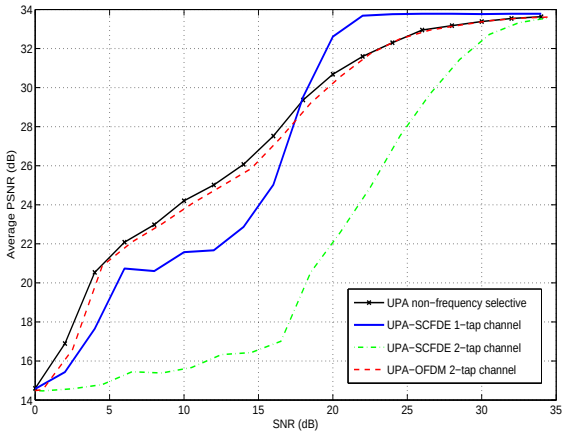


Fig. 4: PSNR performance comparison between combination of the UPA algorithm with SCFDE and OFDM

IV. Simulation Results

We develop our simulation results using Kakadu as the JPEG2000 image coder to encode image of Lena with a size of 512×512 at a rate of 0.25 bit per pixel (bpp). We use Binary Phase Shift Keying as the modulation scheme for the codestream obtained from the codec, and assume that the header information is transmitted error free. Based on our previous experiments in [7], the value of α in (6) is set to be 0.01. We analyze the performance of our proposed UPA scheme with SCFDE technique by calculating the PSNR of the received image transmitted through block fading frequency selective channels. Figure 4 compares the PSNR performance of the system when the UPA algorithm is used alone for image transmission over block fading non-frequency selective channels, and when it is combined with either SCFDE or OFDM to eliminate the effects of ISI and ICI in frequency selective channels. In Fig. 4, a degradation in the PSNR of the received image is noticeable as the number of channel taps increases. For example, at an SNR value of 10 dB, integration of UPA and SCFDE lowers the PSNR of the system in a 2-tap channel for about 6 dB in comparison with the single tap channel. The latter is also lowered by about 2.2 dB compared to the case where only UPA is used for transmission over frequency flat channels. However, this loss in the PSNR performance is not notable when SCFDE is replaced by OFDM. The reason is that our UPA algorithm eliminates the effect of channel to compensate for the instantaneous and average fading. However, SCFDE technique requires the circulant channel structure to obtain multipath diversity and enhance the system performance as the number of channel taps increases. This implies that OFDM is more compatible than SCFDE with our UPA optimization algorithm to combat the negative effect of ISI in frequency selective channels, and maintain high quality of the received image. This figure also suggests that for SNR values greater than 18 dB, the proposed system, which combines UPA and SCFDE, has a superior performance in a single tap channel in comparison to the other scenarios. Figure 5 illustrates the improvement in the PSNR performance when two transmit antennas are employed at the transmitter side using Alamouti's STBC scheme. At an SNR value of 10 dB, this diversity scheme contributes to an increment of about 2.5 dB in the PSNR of the received image transmitted through 2-tap frequency selective channel. Thus, we are able to receive the image with a higher

 ISBN: 978-1-61208-203-5

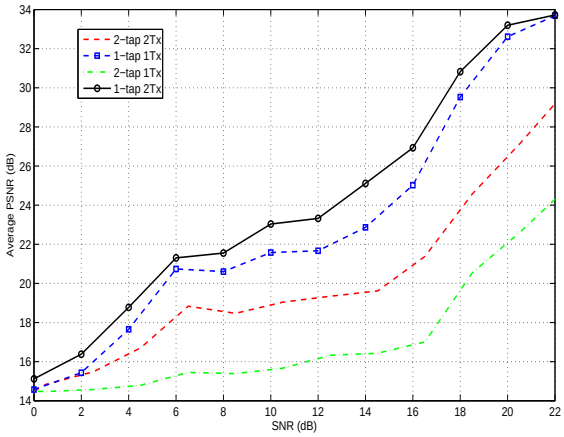
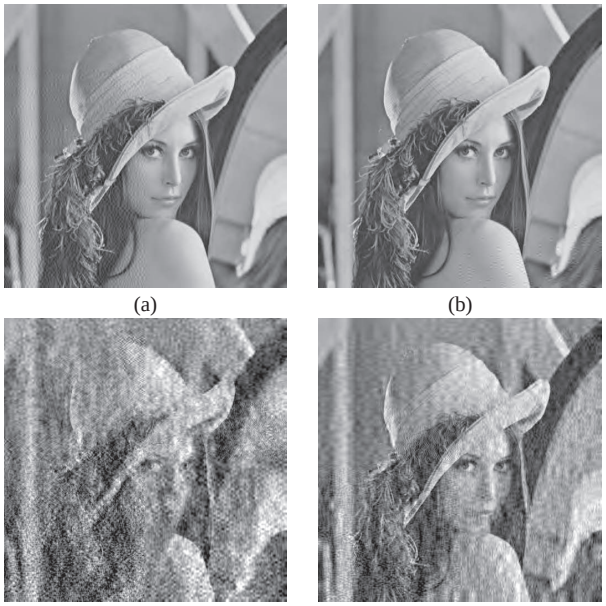


Fig. 5: PSNR performance comparison between different number of transmitter antennas for combination of the UPA algorithm with SCFDE

quality at the receiver, or conserve energy by bounding the quality of the received image.



(a) (b) (c) (d)

Fig.6: Visual comparison of ”Lena” at 0.25 bpp, transmitted at SNR=20 dB over Block fading frequency selective channels (a) 1-tap hannel & 1 transmitter antenna, PSNR=32.70 dB (b) 1-tap channel & 2 transmitter antennas, PSNR=33.74 (c) 2-tap channel & 1 transmitter antenna, PSNR=22.10 (d) 2-tap channel & 2 transmitter antennas PSNR=26.47

Figure 6 presents a visual comparison for transmission of ”Lena” over block fading frequency selective channel using the UPA algorithm and SCFDE technique. The image is transmitted through a single and 2-tap channels with different number of transmitter antennas. It is clear that the visual quality of the received image is enhanced when the number of transmitter antennas increases. In addition, increasing the number of channel taps lowers the quality of the received image.

V. Conclusion and Future Works

In this paper, JPEG2000 images are transmitted through frequency selective block fading channels using an UPA algorithm and SCFDE technique. The optimization algorithm allocates unequal power to each coding pass based on its contribution to the quality of the received image. The simulation results for the image of Lena imply that combination of the UPA algorithm and the OFDM technique leads to a higher image quality than combining the algorithm with the SCFDE technique, while both methods combat the negative effects of ISI and ICI. In this paper, Alamouti’s STBC diversity technique is also incorporated within the proposed system, and the results prove the effectiveness of using two transmit antennas, at the transmitter side, in the PSNR enhancement of the received image. The continuation of this work is to compare the performance of the system at different encoder bit budgets when different number of transmit antennas. This will show us the effectiveness of using two transmit antennas in preserving bandwidth.

References

[1] D.S. Taubman and M.W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practices*, Kluwer Academic. Publishers, 2002.

[2] A. Natu, D.S. Taubman, ”Unequal Protection of JPEG2000 Code-Streams in Wireless Channels,”*IEEE Global Telecom. Conference, GLOBCOM*, vol. 1, pp. 534-538, Nov. 2002.

[3] B.A. Banister, B. Belzer, and T.R. Fischer, ”Robust Image Transmission using JPEG2000 and Turbo-codes,”*Proceedings of the 2000 International Conference on Image Processing*, vol. 1, pp. 375-378, Aug. 2004.

[4] N. Thomas, N. V. Boulgouris, and M. G. Strintzis, ”Optimized Transmission of JPEG2000 Streams over Wireless Channels,”*IEEE Trans. on Image Proc.*, vol. 15, no. 1, pp. 54-67, 2006.

[5] Y. Wei, Z. Sahinoglu, and A. Vetro, ”Energy Efficient JPEG2000 Image Transmission over Wireless Sensor Networks,”*IEEE Global Telecom. Conference, GLOCOM*, vol. 5 , pp. 2738-2743, 2004.

[6] L. Atzori, ”Transmission of JPEG2000 Images over Wireless Channels with Unequal Power Distribution,” *IEEE Trans. on Consumer Electron.*, vol. 49, no. 4., pp. 883-888, 2003.

[7] M. Torki and A. Hajshirmohammadi, ”Unequal Power Allocation for Transmission of JPEG2000 Images over Wireless Channels,” *IEEE Global Telecom. Conference, GLOBCOM*, 2009.

[8] H. Houas, I. Fijalkow, and C. Baras, ”Resource Allocation for the Transmission of Scalable Image on OFDM Systems,”*IEEE Inter. Conference on Comm., ICC*, 2009.

[9] K. Munadi, M. Kurosaki, K. Nishikawa and H. Kiya, ”Robust JPEG2000 Image Transmission over Closed-Loop MIMO-OFDM with Limited Feedback,”*Proceed. of the International Symp. on Circuits and Systems, (ISCAS)*, vol. 2, pp. 432-435, 2003.

[10] Y. Sun, Z. Xiong and X. Wang, ”Progressive Image Transmission over Differentially Space-Time Coded OFDM Systems:Research Articles”*Journal of Wireless Communications and Mobile Computing*, vol. 6, no. 8, Dec. 2006.

[11] U. Sethakaset and S. Sumei, ”Robust JPEG2000 Image Transmission over Closed-Loop MIMO-OFDM with Limited Feedback,”*IEEE Int. Symp. on Personal, Indoor and Mobile Radio Comm. (PIMRC)*, pp. 1-5, 2008.

[12] M. Shayegannia, A. Hajshirmohammadi, S. Muhaidat and M. Torki, ”An OFDM Based System for Transmission of JPEG2000 Images Using Unequal Power Allocation,” accepted at *IEEE Wireless Communications and Networking Conference, WCNC*, April 2012.

[13] F. Pancaldi, G. Vitetta, R. Kalbasi, N. Al-Dhahir, M. Uysal and H. Mheidat, ”Single-Carrier Frequency Domain Equalization, ”*IEEE Signal Processing Magazine*, vol. 25, no. 5, pp. 37-56, 2008.

[14] S. Alamouti, ”A Simple Transmit Diversity Technique for Wireless Communication”*IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451-1548, 1998.

[15] T. Acharya, and P.S. Tsai, *JPEG2000 Standard for Image Compression: Concepts, Algorithems and VLSI Architecture*, John Wiley and Sons, Inc., Hoboken, New Jersey, 2004.

[16] A.H. Sayed, *Adaptive Filters*, Hoboken, New Jersey John & Wiley Sons, Inc., 2008.

[17] N. Al-Dhahir, ”Single Carrier Frequency Domain Equalization for Space-Time Block-Coded Transmissions over Frequency-Selective Fading Channels”*IEEE Communications Letters* , vol. 5, no. 7, pp. 304-306, 2001.

[18] R. Negi and J. Cioffi, ”Pilot Tone Selection for Channel Estimation in a Mobile OFDM System,”*IEEE Trans. on Consumer Electron.*, vol. 44, no. 3, pp. 112-1128, 1998.

[19] C. Shin, R. W. Heath, and E. J. Powers, ”Blind Channel Estimation for MIMO-OFDM Systems,” *IEEE Trans. Veh. Tech.*, vol. 56, no. 2, pp. 670-685, Mar. 2007.

[20] J.G. Proakis, *Digital Communications*, Boston McGraw-Hill, 2000.

 ISBN: 978-1-61208-203-5

Design and Implementation of a Smart Home Energy Management System with Hybrid Sensor Network in Smart Grid Environments

Jinsung Byun, Insung Hong, and Sehyun Park
School of Electrical and Electronics Engineering, Chung-Ang University
Seoul, Korea
E-mail: jinsung, axlrose11421@wm.cau.ac.kr, shpark@cau.ac.kr

***Abstract*— Green IT and smart grid technologies have changed electricity infrastructure more efficiently. Recent advances in wireless and mobile communications technologies facilitate context-aware power management systems which can offer situation-based services in digital home. In this paper, we propose a novel smart home energy management system (SHEMS) with hybrid sensor networks. Hybrid sensor networks consist of two types of sensors: the power information monitoring sensor (PIMS) and the environment information monitoring sensor (EIMS). To maximize the hybrid sensor network lifetime, we propose a new routing protocol based on cooperation between PIMS and EIMS, which we named the CPER. In order to verify the efficiency of our system, we implemented our system in real test bed and conducted some experiments. The results show that the reduction in service response time, the average number of packet transmissions, and energy consumption is approximately 29.8%, 42.3 and 17-22%.**

Keywords- home energy management system; smart Grid; wireless sensor networks (WSNs); pattern-based control, hybrid sensor networks

I. INTRODUCTION

Environmental problems, such as climate change or the exhaustion of natural resources are the one of the most important issues around the world in recent years. These problems are mainly because of the excessive use of energy. To deal with these problems, recently, smart grid technology is emerging and a lot of related works have been done by various researchers and scientists around the world.

Smart grid [1] is defined as a next generation power network that delivers electricity from suppliers to consumers based on two-way communications. This makes it possible for the suppliers and consumers to dynamically respond to changes in energy consumption, demand and grid condition, which improves grid reliability and energy efficiency.

As a part of the smart grid, the micro grid is a low voltage network that interlinks with small distributed power systems. The micro grid provides an independent power grid that interconnects the renewable energy plants with the power storage systems, such as load systems that govern device control in apartments and other dwellings. To enhance the function of micro grids, both energy management systems (EMS) and distributed automation systems (DAS) are needed.

The smart grid [2] designs a smart place which is defined as the energy-efficient place that provides the power-aware and user-centric services according to demand-response (DR) based on an advanced metering infrastructure (AMI) between the users and the power provider. The AMI is the infrastructure which monitors the digital meters, delivers the power consumption information, and controls the various devices. This infrastructure provides the cost and status of the power consumption to the users. The AMI offers an accurate demand forecast to the providers for load management and the revenue protection.

As renewable energy generation and storage systems increase, the power system should manage the demand-response and the power consumption load-balancing with the power storage device. These power systems merge the power provided from the power provider and the power generating from the renewable energy source. The different frequency and voltage are important issue when integrating renewable energy systems into the conventional power networks. Research on distribution and transmission considering integration of the renewable energy system is needed.

Recent studies of energy-aware systems focus on energy monitoring system [3], [4] and energy-savvy device design [5]. Energy monitoring systems [6], [7] allow inhabitants to see energy consumption and control electronic devices to minimize the power consumption of individual appliances. Such systems typically do not consider situation analysis or user satisfaction. Most studies of energy-savvy device design aim to decrease standby power consumption only for specific devices.

To enhance the scalability and effectiveness of power management, existing home network systems and energy-aware systems [3] should consider additional fundamental factors as follows:

 ISBN: 978-1-61208-203-5

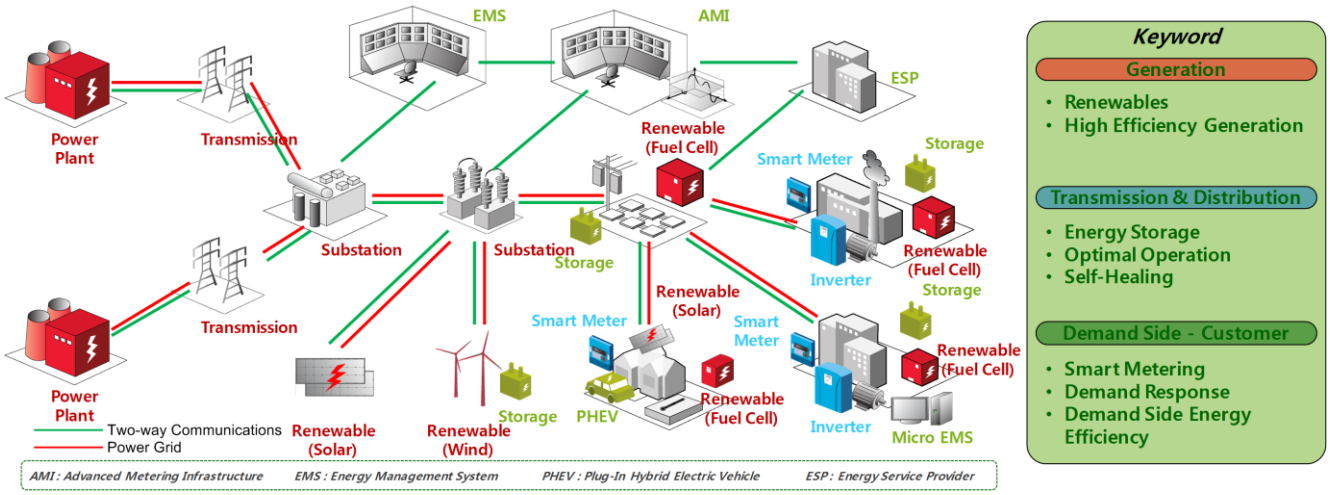


Figure 1. Concept of a smart grid (architecture and keyword)

1) *A deployment of wireless and mobile sensor-dependent environments*: To guarantee QoS and reason adaptive services, existing systems deploy a large number of sensors and then analyze the associated context. For example, crime and disaster prevention services depend on the continuous detection of sensor data. These services do not provide effective power consumption, and do not consider system resources or battery lifetimes, but they do guarantee high-quality service.

2) *The centralized or device-specific scheme*: To analyze contexts and provide service, existing systems use centralized server schemes or service-specific device schemes. However, centralized schemes require increased numbers of sensors to enhance the services that they provide. Moreover, to balancing the demands on system resources this scheme levels the QoS down according to the statuses of devices and sensors. Service-specific devices, which may deliver either static and predefined services are difficult to scale while maintaining appropriate responses to situational changes, such as environmental changes, interruptions, events, movements, or conflicts.

3) *The static rule-based inference*: Existing systems determine the services that they control according to predefined thresholds or on/off schedules. These systems have static policies and service rules with regard to the predefined event analysis and the service determination. Furthermore, due to the high dependency of the device intelligence, the previous systems are less efficient at device control and power consumption in situation management.

Therefore, an energy-efficient system needs to provide effectiveness in power management and user satisfaction for implementing the smart grid. Furthermore, such system must interconnect with intelligent devices, systems, networks, and service provider (SP) by network and information convergence.

Recent work on the smart grid focuses on the load management with AMI and wireless communications infrastructure. The smart grid only collects and monitors the energy status from a home without the consideration of the service management and the power consumption efficiency. Therefore, a power-aware home network system needs to interconnect with the smart grid and manage the energy-efficient services based on the demand-response for implementing the fine grid. Furthermore, the system has to cooperate with the AMI and the renewable storage for the reusability of the smart grid infrastructure.

Considering these requirements of the next generation power grid, we have designed a smart home energy management system (SHEMS). Our system is composed of a smart energy management gateway (SEMG), a smart energy management server (SEMS), and hybrid sensor networks. Our system automatically collects the sensed environmental information and efficiently controls the various consumer devices based on hybrid sensor networks.

II. SMART GRID'S BACKGROUND

A. Network Architecture of Smart Grid

Smart grid basically has the capability to sense power grid conditions, measure power consumption, and control devices with two-way communications. Smart grid makes it possible for the both SP and consumer to dynamically respond to changes in energy consumption, demand and grid condition. Furthermore, reliable and secure access to facilities is crucially important to the success of smart grid. Smart grid is typically composed of three network segments: home/building area networks (H/BANs), AMI or field area networks (FANs), and wide area networks (WANs). Smart grid consists of four parts, which are an electricity generation, transmission, distribution and consumer part (see fig. 1). Smart grid considers improvements in efficiency of all parts (i.e. from a generation part to a consumer part).

The generation part consists of various power plants such as a fossil fuelled power plant, nuclear power plant, hydroelectricity plant, and renewable power plant. Safety and reliability are most important factors in a generation part of smart grid. In smart grid, information about the

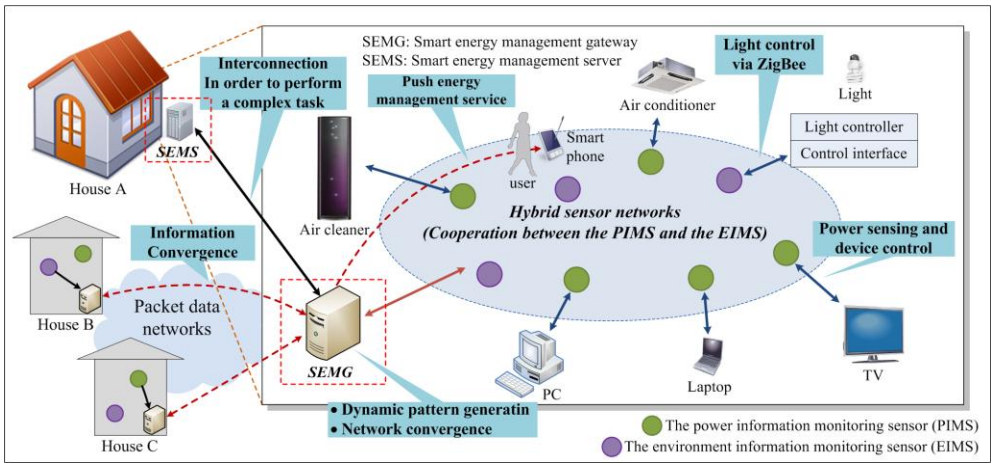


Figure. 2. Overview of the smart home energy management system

condition of the plant and environment is transmitted to administrators via IT infrastructure, thereby making the power plants more safe and reliable. Furthermore, smart grid is able to synchronize and adjust the voltage output of the added generation without damaging the whole system.

The transmission and distribution part play a role in optimizing the electricity transmission and distribution. That is, the loss and cost during transmission and distribution have to be minimizedm Alternating current (AC) is typically preferred since its voltage easily amplifies by a transformer, which minimizes resistive loss in the conductors used to transmit electricity over long distances. Recently, high-voltage direct current (HVDC) is used to deliver electricity due to the advantage of which is to transmit large amounts of electricity over long distances with lower costs and losses. In addition, key technologies enabling transmission and distribution part deployment are distribution automation system (DAS), supervisory control and data acquisition (SCADA), robust faulty detection, self-healing, and so on.

The consumer part is an end user (or consumer) of electricity and consists of many types of consumers such as home, building, and industry. Main role of this part is to gather and transmit various contexts such as power consumption or state of power system. Furthermore, this part manages electric vehicle (EV) charging and local renewable energy sources. In addition, because the smart grid's goal is enhancement of energy efficiency through response to many conditions in supply and demand, the smart meters are deployed to gather various contexts which are applied to calculate the supply and demand.

B. Communication Standardization for Smart Grid

There are some short and medium range wireless communications technologies emerged in the field of smart grid.

1) *IEEE 802.11 (WiFi)*: The IEEE 802.11 protocol is a set of standards for implementing a wireless local area network (WLAN), which is suited for high-data-rate applications over large areas. The IEEE 802.11 standard, commonly referred to as WiFi, is the most accepted technology for indoor wireless communications. The basic shortcoming of this standard is the high power requirement of devices.

2) *IEEE 802.15.1 (Bluetooth)*: IEEE 802.15.1 standard, commonly referred to as Bluetooth, is an open wireless technology standard for transmitting and receiving data over small areas. Bluetooth utilizes short wavelength radio transmissions in the ISM band from 2400 to 2480 MHz, creating personal area networks (PANs) with high levels of security. The Bluetooth protocol is well suited for low-power/low-data-rate applications. The main shortcoming of this standard is scalability. That is, Bluetooth networks support up to only eight devices. Another shortcoming is its periodical waking up and synchronization with the master device of Bluetooth networks.

3) *IEEE 802.15.3 (UWB)*: IEEE 802.15.3 is a MAC and PHY standard for high-rate WPANs. There are two major types of application that uses UWB communication. The first type of application is for high-data-rate (over 1 Mb/s) communications such as multimedia content transmission. The other type of application is for low-data-rate (below 1 Mb/s) such as wireless sensor networks. The main shortcoming of this technology is similar to that of WiFi. That is, the high power requirement.

4) IEEE 802.15.4 (ZigBee): ZigBee is a standard which is employed in many home networking solutions because ZigBee has a low-power and low-cost characteristics. The ZigBee provides various network topologies such as a cluster tree, self-healing mesh network, or star topology, according to application's requirements. In this way, the fixed and mobile devices can be configured flexibly. In addition, a ZigBee device using carrier sense multiple access with collision avoidance (CSMA/CA) does not require scheduling special wake-up events in order to communicate and maintain synchronization. Thus, ZigBee presents itself as a much better candidate (for wireless communication in the HAN) than UWB, WiFi, and Bluetooth [8].

 ISBN: 978-1-61208-203-5

III. SYSTEM ARCHITECTURE

A. Overview of SHEMS

In this section, we present the overall system architecture of SHEMS. Fig. 2 shows an overview of the proposed smart building energy management system. The proposed system consists of hybrid sensor networks (the power information monitoring sensor + the environment information monitoring sensor), the smart energy management gateway (SEMG), and the smart energy management server (SEMS). We present our system in more detail below:

- The Hybrid Sensor Networks: To establish the proposed hybrid sensor networks, we utilize an intelligent sensor that is used for sensing the context as well as controlling the device according to rules or policies. For example, the environmental information monitoring sensor (EIMS) basically plays a role of information (e.g. the temperature, the humidity, the intensity of illumination, etc.) sensing and transmission to SEMG. However, it also directly controls a TV, a fan, and a light. In addition, the hybrid sensor networks consist of the two types of sensors: the power information monitoring sensor (PIMS) and EIMS. These sensors operate based on the proposed routing algorithm, the CPER.
- SEMG: SEMG can efficiently distribute various tasks related to the energy management service based on the hybrid sensor networks. It also interoperates with various mobile devices, such as smart phone, PDA, notebook PC using IEEE 802.15.4 (ZigBee), IEEE 802.11 (WLAN) technology.
- SEMS: SEMS performs complex tasks, such as load forecasting, user and device authentication /authorization, complex events analysis, etc. SEMS also manages the whole building energy and environmental information, user and device profiles. In addition, it performs a task of interconnection with other SEMSs.

B. Hybrid Sensor Network

In this subsection, we address the hybrid sensor networks architecture used for the proposed home energy management system. Our hybrid sensor networks consist of many smart sensor nodes. They can perform several tasks, such as gathering real-time energy consumption and building environmental information as well as controlling the various consumer devices. The proposed hybrid sensor networks are divided into two groups: PIMS group and EIMS group.

1) *PIMS*: This is mainly used for gathering the power consumption and the power state of the consumer device directly connected to PIMS. It is also used to directly control the consumer device. It has ZigBee and power line (PL) based communication capability in order to automatically establish sensor networks. The information about the power consumption and the power state collected by PIMS is transmitted to the SEMG. The SEMG analyzes this information and then generates a power consumption pattern. If an energy management operation is needed, the SEMG sends the control signal to PIMS in order to control and manage the consumer device. PIMS broadcasts its sensor identifier (SID) periodically once it enters the local hybrid network, so that the SEMG can recognize PIMS.

2) *EIMS*: This sensor monitors the environmental information, such as the temperature, the humidity, the gas (LPG and LNG), carbon monoxide (CO), the intensity of the illumination, the user's movement, etc. This sensor also generates a device control signal and directly transmits it to PIMS in order to control the consumer device. Like PIMS, EIMS has ZigBee-based communication capability in order to establish wireless sensor networks.

3) *CPER - Cooperation between the PIMS and EIMS based Routing protocol*

The main goal of the CPER protocol is to increase the lifetime of the hybrid sensor networks through the cooperation of the two types of sensors (PIMS and EIMS). The important difference between the two types of sensors is that one operates using the electrical power from the power socket directly and the other operates using that from a finite battery. We utilize these properties to route a packet from the source to the destination.

A number of existing wireless sensor network routing protocols, such as the ones found in [9] and in [10], do not consider an adaptive allocation of the system resources and the user-centric service aspects; they are not adequate in ubiquitous environments where novel services are provided for various users. Therefore, we propose a new routing protocol that is cooperative method between PIMS and EIMS, which is called CPER. We design our routing protocol suitable for home energy management services. The proposed protocol establishes the wired and the wireless sensor networks, based on cooperation between the two types of sensors in order to maximize the network lifetime. The protocol utilizes difference between the two types of sensors as mentioned above. The CPER protocol works as follows:

* Clustering: The SEMG determines a cluster-head depending on the node's power supply types. That is, the SEMG initially elects a node with a direct power source as a cluster-head. In the case of a node with battery-powered node, the SEMG applies the simplified LEACH [11] -based protocol to the cluster-head selection.

* Route discovery: (1) (default) If the source should discover a route to the destination, it broadcasts a route request packet (ROUTE_REQ) to its neighbors. If a node receiving a ROUTE_REQ does not know a route to the destination, the node inserts its own address into route tracking field of the packet and transmits a modified packet to its neighbors. In this way, paths are tracked. A typical problem for many ad hoc routing protocols is the needless packet flooding. To avoid the unnecessary packet flooding (e.g., endless packet circulation), each node only forward packets it has not yet seen. In addition, a ROUTE_REQ carries a form of expiration information, such as maximum number of hops in order to avoid unnecessary packet transmission. When receiving ROUTE_REQs that have different route, it selects a ROUTE_REQ with the minimum-

 ISBN: 978-1-61208-203-5

Figure. 3. Prototype of the EIMS and the PIMS

hop path and sends a route reply a packet (ROUTE_RE) to the source node. (2) (cooperation with PIMS) If a node receiving a ROUTE_REQ has PIMS as its neighbor; does not know a route to the destination, it adds its own address to the packet and transmits (unicast) the ROUTE_REQ to PIMS. Long-distance transmission causes the significant energy consumption due to the attenuation characteristics of radio-frequency (RF) signals. Therefore, we use the battery-operated EIMS for short-distance transmission and use PIMS for relatively long-distance transmission. In addition, because PIMS also have the wired communication (i.e., PLC) capability, this makes the service response time faster.

* Data forwarding: Considering the requirements of network architecture for a home energy management service, we use the next-hop routing scheme. This approach supports some level of fault tolerance and makes our system more robust and resilient to node mobility.

4) *Implementation*: The hybrid sensor networks consist of two types of sensors: PIMS and EIMS (see fig 3). We designed each sensor node with sensing, controlling and networking abilities. Each sensor node automatically establishes the hybrid sensor network using proposed scheme and protocol. We designed PIMS using the 8-bit microprocessor, and a ZigBee transceiver for communication with other sensors and the SEMG. We use the ZigBee technology due to its low-cost and low-power characteristics [12]. PIMS also has a power line communication (PLC) module to communicate with the SEMG and other PIMSs. A wireless link has many advantages, such as high scalability and easy deployment compared with a wired link. However, since a wireless link also has a number of drawbacks, such as decreasing signal strength, frequent signal collision, and

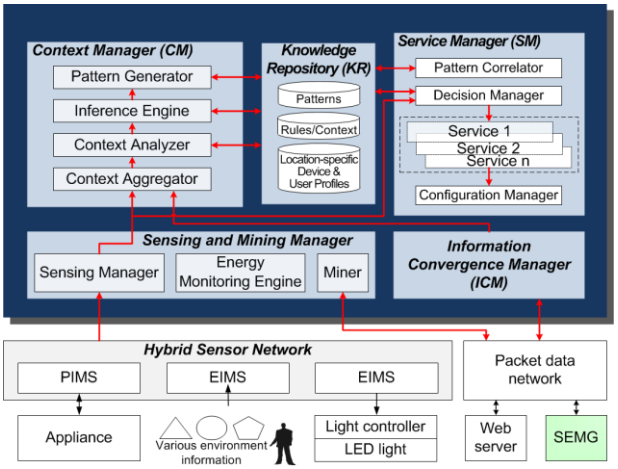


Figure 4. Middleware architecture of SEMG

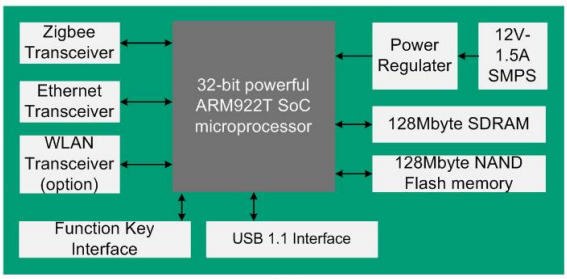
fading, our system balances between the wireless link and the wired link. EIMS is equipped with a low power 8-bit microprocessor and a ZigBee module for communication with other sensors and the SEMG, much like PIMS. EIMS analyzes the user's situation and surroundings and operates the rule-based engine through the main processor group composed of the low-power 8-bit microprocessor and memory. EIMS has various sensor modules, such as temperature, humidity, gas detection (LPG, LNG, CO), Carbon dioxide, and object detection.

C. *Architecture of SEMG*

1) *Middleware Architecture*: Fig. 4 shows middleware architecture of the SEMG. It consists of various components. We will present the core modules in more detail.

- *Context Manager*: It gathers the sensor data and categorizes the situational events for the classification and storage to *Knowledge Repository*. This module assorts the meaning events that their values have the effects of the service status or the conflict situation. It transmits the meaning events to Inference Engine and requests the pattern verification to *Pattern Correlator*.
- *Inference Engine*: In order to reasoning and predicting the adaptive service, this module validates

(a)



(b)

Figure 5. (a) Prototype and (b) hardware blockdiabram of the SEMG

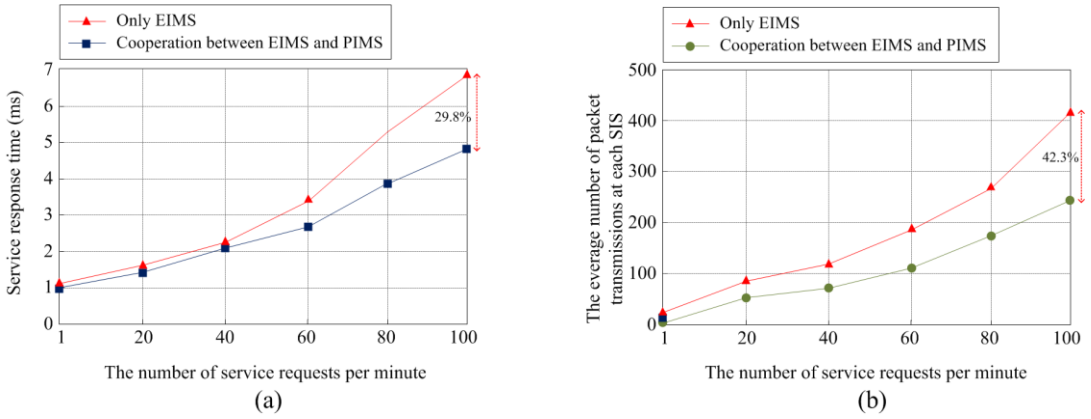


Figure. 6. Comparison of (a) the service response time, and (b) the everage number of packet transmissions at each EIMS

the correlation the events with the service pattern and analyzes the contexts receiving form *Knowledge Repository*. It decides whether or not to maintain the service and conflict solution to *Service Manager* with the pattern and policy.

- *Service Manager*: This module monitors the service status and maintains the personalized service. Moreover, this is used to determine the transmission of the service status according to the domain interconnection and correlation policy.
- *Configuration Manager*: For the convergence and the interconnection, this module manages the heterogeneous contexts which consist of the status of intelligent device, the system, and the network in a scalable fashion. When the environmental elements change in various situations, this module transmits the control signal that reconfigures the address and status.
- *Energy Monitoring Engine*: This module monitors the total and moment power consumption. Thus, this analyzes the pattern of consumption and verifies the correlation with service. When the power status suddenly changes or made up the novel pattern, this module interworking with *Inference Engine* actively organizes the differential service.
- *Pattern Correlator*: For the convergence of the network, the device, the service, and the system resource, this module generates the patterns and analyzes the correlations. In addition, this module influences the policy modification and the service prediction based on efficiency QoS requirements.

2) *Implementation*: Fig. 5 shows the prototype and hardware block diagram of SEMG. The main processor is based on 32-bit powerful ARM922T SoC (System on Chip) microprocessor. It is used for analyzing the complex events, operating the middleware, and processing the pattern generation. It also controls PIMS and EIMS. The communication group consists of a ZigBee transceiver, and Ethernet and WLAN modems. We used a 250kbps/2.4GHz ZigBee transceiver and 10/100Mbyte Ethernet modem for communication. A smart phone or a smart pad using IEEE 802.11 (WLAN) can receive the energy management service via a WLAN modem. The electricity monitoring group plays a part in monitoring the power consumption and the power state. Furthermore, when electricity leakage or overvoltage happens, the SEMG recognizes abnormal events and autonomously takes steps to counteract or alleviate these problems. The power group is composed of a SMPS and a power regulator.

We also implemented the SEMS to process various complex tasks (e.g. electric load forecasting, three-dimension (3D) simulation based on the energy management), a user and device authentication/authorization). There are various electric load forecasting methods and techniques, such as neural network based method [13] and time-series based model [14]. We adopted an ARIMA model, because in theory, it is the most typical model and possible to easy implementation.

IV. EXPERIMENT AND RESULT

Fig. 6 (a) shows the service response time by the number of information requests per minute. Even though the number of service requests increases, our system using CPER protocol maintains certain levels of delay of the request and response. Our system also gradually decreases the slope of the service response time according to a new routing

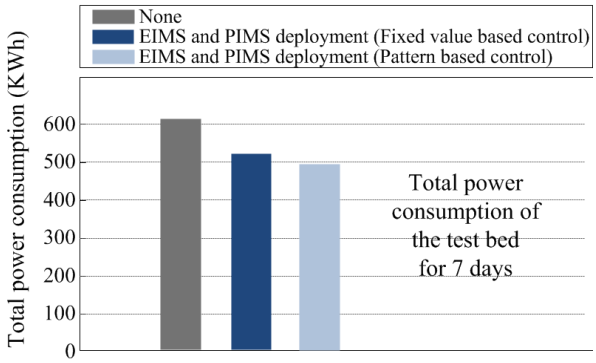


Figure 7. total power consumption of the test bed

protocol based on cooperation between PIMS and EIMS, whereas not cooperating with PIMS when routing rapidly increases the service response time due to the frequent packet collision and packet loss. The results show that the service response time reduction using our system is approximately 29.8% under conditions for generating 100 service requests per minute. Fig. 6 (b) presents the average number of packet transmissions at each EIMS by the number of information requests per minute. Similar to the result shown in Fig. 6 (a), proposed system using CPER protocol gradually decreases the slope of the average number of packet transmissions at each EIMS. The results show that the reduction of the every number of packet transmissions using our system is approximately 42.3% under conditions for generating 100 service requests per minute. Because the battery power consumption at a sensor node is proportionate to the number of packet transmissions, we can increase the network lifetime through our hybrid sensor networks. Fig. 7 illustrates the results of total power consumption for 7 days. The results show that the power saving using our system is approximately 17-22% by utilizing our energy management services, such as light control by using sensing data from EIMS, shutting up the standby power, the remote power control using a smart phone, etc.

V. CONCLUSION

Green IT technology is emerging and many related works have been done by various researchers around the world. In this paper, we propose a smart home energy management system (SHEMS) architecture based on hybrid sensor networks to make consumer devices more energy efficient and intelligent. We also present a new routing protocol to increase the hybrid sensor networks lifetime. We named this routing protocol the CPER, whose basic idea is in cooperation between the power information monitoring sensor (PIMS) and the environment information monitoring sensor (EIMS). We implemented our system, and we design and develop related hardware and software. We expect that our work will contribute to the development of novel home energy management system. In order to verify the efficiency of our system, we implemented our system in real test bed and conducted some experiments. The results show that the reduction in service response time, the average number of packet transmissions, and energy consumption is approximately 29.8%, 42.3 and 17-22%.

As a part of our future works, we are doing research into novel context awareness technologies and trying to apply them into an energy management system. Furthermore, we are developing a self-organized energy management system for various environments.

ACKNOWLEDGMENT

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the HNRC(Home Network Research Center) –ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency (NIPA-2010-C1090-1011–0010) and by the Human Resources Development of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea government Ministry of Knowledge Economy (20104010100570).

REFERENCES

[1] Z. Pei, L. Fangxing, and N. Bhatt, "Next-Generation Monitoring, Analysis, and Control for the Future Smart Control Center," *Smart Grid, IEEE Transactions on*, vol. 1, pp. 186-192, 2010.

[2] H. Farhangi, "The path of the smart grid," *Power and Energy Magazine, IEEE*, vol. 8, pp. 18-28.

[3] L. Chia-Hung, B. Ying-Wen, and L. Ming-Bo, "Remote-Controllable Power Outlet System for Home Power Management," *Consumer Electronics, IEEE Transactions on*, vol. 53, pp. 1634-1641, 2007.

[4] S. Darby, "The effectiveness of feedback on energy consumption: a review for DEFRA of the literature on metering, billing and direct displays," Environmental Change Institute, University of Oxford 2006.

[5] H. Joon, H. Choong Seon, K. Seok Bong, and J. Sang Soo, "Design and Implementation of Control Mechanism for Standby Power Reduction," *Consumer Electronics, IEEE Transactions on*, vol. 54, pp. 179-185, 2008.

[6] Kurt Roth and J. Brodrick, "Home Energy Displays," *ASHRAE Journal*, vol. 50, pp. 136-138, 2008.

[7] L. F. Stein, "California Information Display Pilot: Technology Assessment," 2004.

[8] Z. M. Fadlullah, M. M. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, and Y. Nozaki, "Toward intelligent machine-to-machine communications in smart grid," *Communications Magazine, IEEE*, vol. 49, pp. 60-65, 2011.

[9] S. D. Muruganathan, D. C. F. Ma, R. I. Bhasin, and A. O. Fapojuwo, "A centralized energy-efficient routing protocol for wireless sensor networks," *Communications Magazine, IEEE*, vol. 43, pp. S8-13, 2005.

[10] W. Chen, Z. Chen, F. Pingyi, and K. Ben Letaief, "AsOR: an energy efficient multi-hop opportunistic routing protocol for wireless sensor networks over Rayleigh fading channels," *Wireless Communications, IEEE Transactions on*, vol. 8, pp. 2452-2463, 2009.

[11] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *Wireless Communications, IEEE Transactions on*, vol. 1, pp. 660-670, 2002.

[12] A. Wheeler, "Commercial Applications of Wireless Sensor Networks Using ZigBee," *Communications Magazine, IEEE*, vol. 45, pp. 70-77, 2007.

[13] C. Ying, P. B. Luh, G. Che, Z. Yige, L. D. Michel, M. A. Coolbeth, P. B. Friedland, and S. J. Rourke, "Short-Term Load Forecasting: Similar Day-Based Wavelet Neural Networks," *Power Systems, IEEE Transactions on*, vol. 25, pp. 322-330, 2010.

[14] E. Gonzalez-Romera, M. A. Jaramillo-Moran, and D. Carmona-Fernandez, "Monthly Electric Energy Demand Forecasting Based on Trend Extraction," *Power Systems, IEEE Transactions on*, vol. 21, pp. 1946-1953, 2006.

 ISBN: 978-1-61208-203-5

Research on improving accuracy of Cardiac Disorder data analysis based on Random Forest classifier

HyunJu Lee[1], DongIl Shin[1] and Dongkyoo Shin[1]
Department of Computer Engineering
Sejong University
1. e-mail: nedkelly@gce.sejong.ac.kr,
{dshin, shindk}@sejong.ac.kr

HeeWon Park[2] and SooHan Kim[2]
Visual Display Div. R&D Team
SAMSUNG Electronics Co. HQ
2. e-mail:{heewonpark, ksoohan}@samsung.com

***Abstract*— In order to prove that the improved RF algorithm had higher accuracy, the comparing analysis was conducted adapting ECG data. In pre-processing stage, Band-pass Filter was adapted among Wavelet transform, Median Filter, Finite impulse response and others. As a result, the modified Random Forest classifier showed increased more accuracy than SVM, MLP and other researchers' results. Thus, continuous studies on the selection of the filters and methods, which can efficiently delete baseline-wandering at pre-processing phase and accurately extract R-R interval, should be taken place.**

Keywords-ECG; R-R interval; HRV; SVM; MLP; Random Forest; classifier; accuracy.

I. INTRODUCTION

ECG (Electrocardiogram) is an electric signals released by heart activities, which is used as a reference that can identify conditions and diseases of the heart [1]. ECG consists of five ripple marks; P, Q, R, S and T, which verify signals according to height of ripple marks and features of interval, and also can compose ECG data through decision making whether disease exist or not. There is arrhythmia which can be detected by ECG signals, which generally means irregularly fast and slow blood beats [2]. There is MIT-BIH Arrhythmia Database which published for research on arrhythmia.

Signals of ECG are generally experimented based on R-R interval and QRS-Complex extracted data from ECG. Tsipouras, Fotiadis, and Siderise [3] detected and classified arrhythmia according to generated features of heart beat from R-R interval signals. Firstly, they detected signals with blood beats verifying from arrhythmia signals, and then, arrhythmia extraction tasks were secondly conducted with six features released from arrhythmia signals. SVM and MLP classifiers are the most frequently used on ECG experiments. Asl [4], who experimented HRV, proceeded the experiment by two ways; GDA (Generalized Discriminant Analysis) method which is Dimension reducing method was applied into one case of the experiment and GDA was not adapted in another case.

However, it is necessary that experiments on the performance of Random Forest classifier which has differing algorithm compared to SVM and MLP are needed to improve accuracy on experiment results in arrhythmia. Thus, in this study, comparative analysis on accuracies between SVM and MLP classifier was conducted to find out performance of Random Forest classifier. In addition, comparative analysis between parallel data of other researches which experimented with R-R interval extracting and results of this study was also undertaken. R-R interval signal data were verified and constructed, drawing on beat annotation provided by MIT-BIH Arrhythmia Database, and also, modifications of classifier algorithm were attempted.

In this study, there are three different contents in each paragraph state below:

The explanation related to data as well as the process of the experiments was represented in the Section 2. Then, the explanation of the algorithm and the results were commonly noticed in the Section 3. Finally, the conclusion of this study and the direction of further researches were recorded in the Section 4.

II. RELATED WORKS

Meanwhile, there were a lot of experiment concerned with ECG signals and have been applied various filters and classification algorithms. In the case of filters, there were Chazal's [5], Michael's [6], Martinez's [7] works, and so on. Chazal experimented with median filter [5], Michael tested with FIR (finite impulse response) [6], and Martinez tested with wavelet transform [7], however, we experimented with the band-pass filter like Markovsky [8], Taouli [9], and Gholam-Hosseini [10], the band pass filter was judged to be superior to the others and efficient to distinguish the wavelets of ECG by separating whether narrow or wide wavelet.

In the case of classification algorithms, most of which were generally SVM (Support Vector Machine), MLP (Multilayer Perceptron), and DT (Decision Tree), Chau [11], Asl [4], and Bsoul et. al. [12] experimented with SVM and Zhang [13] with the combination of PCA (Principal Characteristics Analysis) and SVM. Also, Inan [14], Yaghouby [15], and Ozbay [16] tested with MLP and Quinlan [17] and Exarchos [18] with DT. And also, Mahesh [19] experimented with Random Forest, Logistic Model Tree, and MLP in classifying the cardiac diseases. Now-a-day, it has come up to more than 90% of accuracy in classifying ECG signals, This paper try to research another method to obtain more accurate rate of classification than existing ones by using the revised Random Forest classifier.

 ISBN: 978-1-61208-203-5

III. DATA AND PRE-PROCESSING

A. *MIT-BIH Arrhythmia Database*

MIT-BIH (The Massachusetts Institute of Technology – Beth Israel Hospital) Arrhythmia Database [20] is a researched data related arrhythmia analysis with supports receiving from Boston's Israel Hospital and MIT since 1975. MIT-BIH Arrhythmia Database is the first arrhythmia data which can be universally used to detect and evaluate arrhythmia, and total data records are digitalized records from 360 samples per hour per channel. It is ECG records which had been researched in BIH arrhythmia laboratory between 1975 and 1979, measuring patients' movements such as walking through two channels during 24 hours. The database consist of 48 data: 23 numbers of records which were randomly collected from recorded 4000data sets were selected from 40% of outside patients and 60% of hospitalized patients. And other 25 numbers of data included significant arrhythmia signals in clinic although the data were collected from the same patients group.

B. *Feature extraction of R-R interval*

R-R interval means time of R wave in a human's brain from one certain peak to a next peak, and each R-R interval consists of one cardiac cycle. Fig. 1 indicates R-R interval [21].

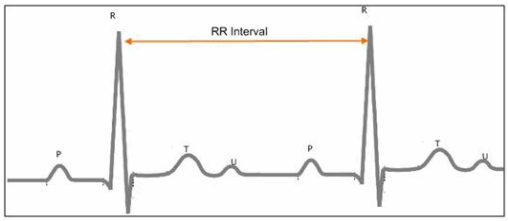


Figure 1. A sample image of R-R interval

R-R interval is continuously generated as a form of continuous time, which is repeated. Sequence of R-R interval are transformed when QRS detector is applied in to ECG signals [22].

Sequences of R-R interval are constituted through time succession, and each sequence which corresponds to immediate heart proportion is defined by the below formula [22].

$$F_i = 1 / RR_i \quad (1)$$

In general, HRV analyzes HRV in extracted R-R interval using HRV Analysis and constructs HRV data based on analysis information of the extracted HRV. HRV is distinguished into below properties Mean, RMSSD, SDNN, SDSD, NN50, pNN10, pNN5 and so on. In this study, the data properties were classified into total 25 categories including Mean, RMSSD, SDNN, pNN50 and others.

- Mean: inquiring meaning of the 32 number of R-R interval values in each segment.
- RMSSD: meaning the average value of RMS (Root Mean Square) among gaps of intervals from R-R interval.
- SDNN: meaning standard deviation of the gap of R-R interval.
- pNN50: meaning proportions from total section in cases that the gap of R-R interval is over 50cm.

Fig. 2 indicates the feature extraction of R-R interval, and Fig. 3 illustrate HRV analysis. The filter is not only used to delete unnecessary components (frequency components), but also exchange measured data; distances, speeds, accelerations, temperature and strengths, into electric signals. For example, there are Median Filter, finite impulse response, Wavelet transform, Fourier transform and Band-pass Filter, which function as the device (stated above).

In this study's experiments, since Biomedical Startup Kit 3.0 provided by NI LABVIEW (National Instrument LABVIEW) was applied in extraction tasks, Band-pass Filter provided by the kit was adopted. (Fig. 4) Band-pass Filter was designed to filter noises with combining low-pass and high-pass in a single filter [23].

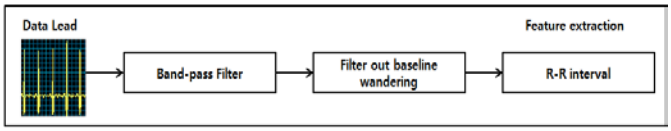


Figure 2. Feature extracion in R-R interval

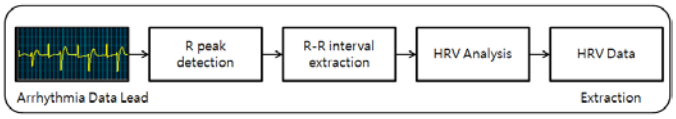


Figure 3. HRV analysis

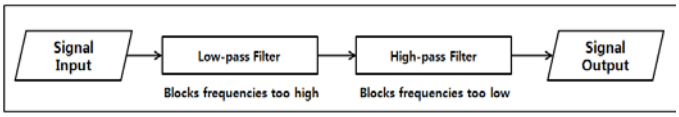


Figure 4. Input and output modes of Band-pass Filter

High-pass and low-pass of the filter were configured at 25Hz and 10Hz respectively. Configured filters erased noises of the data signals, and extracted R-R interval through deleted signals. Then, R-R interval was designed a form to be experimented by WEKA which was used in the classifier experiment. Finally, designed data were experimented by Random Forest [25] classifier, which was one of the classifiers provided by WEKA. Fig. 5 and 6 indicate arrhythmia data before feature extraction and after the extraction. Extracted signals were classified into normal signals and arrhythmia signals according to their intervals and heights. RF is an algorithm belonged to 'tree'. The accuracy of RF was reinforced compared to AF (Atiral Fibrillation) in [27] which had been compared in this study.

 ISBN: 978-1-61208-203-5

And the experiment was performed in [4] with SVM and MLP algorithms applied. In terms of the accuracy of the result, RF relatively showed a higher performance. Therefore, the experiments were undertaken based on RF and the algorithm was also modified to improve the accuracy in this study.

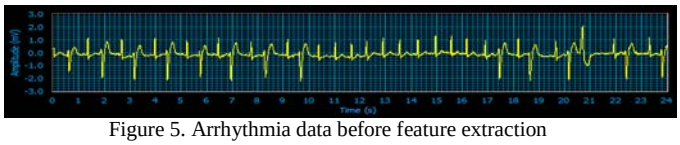


Figure 5. Arrhythmia data before feature extraction

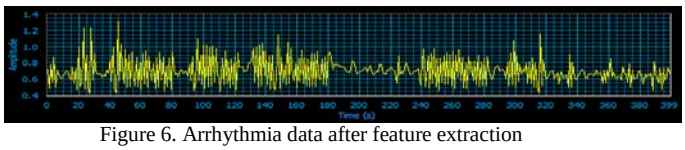


Figure 6. Arrhythmia data after feature extraction

IV. BODY

A. *Modified algorithm of Random Forest classifier and formula*

1) *Modified algorithm of Random Forest*

In 1998, through a research, Ho noticed that Random Subspace is a method to select randomly from each tree with grown subsets using. And after a year, Breiman used new analysis data which was designed to extract results randomly in the original analysis data [24]. Random Forest which is an algorithm that selects random vectors is a specially designed ensemble technique for Decision Tree classifier [25]. Each Decision Tree uses random vectors created from certain possibility distributions. When the tree grow, Decision Tree defines random vectors to segregate each node from selected input features of F numbers rather than totally investigates input features of F numbers selected randomly [25]. It has a input feature called Forest-RI and Forest-RC: Forest-RI is a way which randomly select a vector of the RI, Forest-RC divides input data into the beat condition when input features of F numbers reach universal linear compounding [25]. In modified algorithm in this study, Forest-RI was designed to select the most frequent signals and Forest-RC was designed to classify arrhythmia chased by the algorithm. And Best-First decision tree (B-F tree) was applied rather than Decision Tree.

In general, Decision Tree which classifies target variables has had an aim that classified a given data. Also, selecting the most related variables and target variables, tree compounds categories and separates the most related category, which tree has a limitation according to features of the basic data.

Thus, tree cannot guarantee the best accuracy because tree becomes too sophisticated and the rate of the classificati on shows low performance when the features of the data are not vertically classified to certain variables. Therefore, to complement these drawbacks, B-F Tree is applied to modify algorithm. B-F Tree which is a method that extends nodes with best-fit order rather than fixed orders minimizes errors which come from all nodes needing separation with the most efficiently separated nodes added in each experiment stage. In each stage, tree extends with the most modified subset selecting. And the constructed process is expanded when all of the nodes reach a certain number or a pure node [26]. At a stage of pruning, the first B-F Tree can conduct two methods; pre-pruning and post-pruning.

When tree is growing, pre-pruning stops its growth, if data, which are divided, are not practical. The second post-pruning, which continuously extends nodes until all of the trees are completely extended, and it selects with extended data numbers and sorts of the average error estimates-minimizing [26]. Two cases were made based on all data of the final tree and extended selecting numbers.

2) *Formula*

Through experimenting Random Forest classifier, Accuracy, Sensitivity, Specificity and PPV (Positive Predictive Value) were measured after TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative) had been extracted. The formula is stated below.

$$\text{Accuracy} = TP + TN / TP + TN + FP + FN \quad (2)$$

$$\text{Sensitivity} = TP / TP + FN \quad (3)$$

$$\text{Specificity} = TN / TN + FP \quad (4)$$

$$PPV = TP / TP + FP \quad (5)$$

If results of the formulas are needed to exchange into percentage, each results of the formulas is just multiplied 100.

B. *The method of the experiment*

In this study's experiments, R-R interval was extracted as two wave forms; Narrow and Wide with Band-pass filter using. In constructed arrhythmia data, N (Normal) and ~ (Change in signal quality) signals were the most frequent beats: N means the regular wave form and ~ reveals that wave forms become to shift their current form. V (Premature Ventricular Construction) and A (Atrial Premature Beat) signals, which are commonly mean arrhythmia, had been generated the most frequently in this study's experiments, therefore, the algorithm was modified based on those the most frequent beats. In the modified algorithm, Forest-RI preferentially chose N (Normal) and ~ (Change in signal quality) signals, and Forest-RC were designed to classify V (Premature Ventricular Constraction) and A (Atrial Premature Beat) signals after chasing them. And then, after N and ~ signals were separated, Forest-RI was constructed to chase and classify other signals as well. The modified Random Forest algorithm is stated below.

 ISBN: 978-1-61208-203-5

- Forest-RI firstly chases F = N (Normal) and F = ~ (Change in signal quality), then verifies them.
  - N = Normal signal
  - ~ = altering signals // Forest-RI means input selection
- Forest-RI chases V (Premature Ventricular Constraction) and A (Atrial Premature Beat)
- Forest-RC distributes V and A into arrhythmia // Forest-RC means the highest separation
  - V = Arrhythmia
  - A = Arrhythmia
- Forest-RI chases other signals and distributes, excepting N, ~, V and A signals
- Forest-RI and Forest-RC are repeated
- Forest-RI / Forest-RC are ended

In this study, apart from the modification of Forest-RI and Forest-RC, Beat-First decision tree (B-F Tree) was applied rather than Decision Tree in order to reduce out-of-bag. B-F Tree, which uses best-fit order to extend nodes without fixed orders, stops its growth otherwise segregated data do not show actuality, and it makes decisions with finally extended data volumes adapting. Thus, it can decrease out-of-bag rates more easily than Decision Tree as minimizing extended volumes and branches. When Forest-RI and Forest-RC had been able to chase and classify selectively, TP (True Positive) showed relatively high values before its modification, while values of FP (False Positive) were decreased. And out-of-bag was relatively reduced compared to the past experiments when B-F Tree had been applied. Therefore, its accuracy remarkably higher than the results of other established experiments. A selected datum was experimented to identify that the performance of B-F Tree was more excellent than Decision Tree. From the result, two facts stated below were revealed. The accuracy of Decision Tree was 90.69%: its volume and leave were 341 and 171 respectively, whereas the accuracy of B-F Tree was 93.37% as its volume and leave of tree were 567 and 284 respectively. This study's actual classifier experiments were conducted with Random Forest classifier in WEKA–3.6.2 version Fig. 7. In the experiments, R-R interval had been extracted at the pre-processing stage, and extracted data were then corrected to be experimented by WEKA. Experiments using WEKA had been firstly compressed, Random Forest classifier experimented the compressed data. The experiment using Random Forest was undertaken by k-fold-cross-validation method. Seperating data as k number of times of the same size section, k-fold-cross-validation method is a means that, an experimental section is selected among other sections and the others are used as training materials [25]. According to these sequences, each section is repeated in order to be used exactly once only, and total out-of-bags added by k number of times of the total experiments. In this study, configuring k to 10, experiments had been performed by using 10-fold-cross-validation, and then, Accuracy, Sensitivity, Specificity and PPV (Positive Predictable Value) were measured based on extracted figures of TP, TN, FP and FN. And Accuracy, Sensitivity and Specificity were just measured from feature data of HRV in this study.

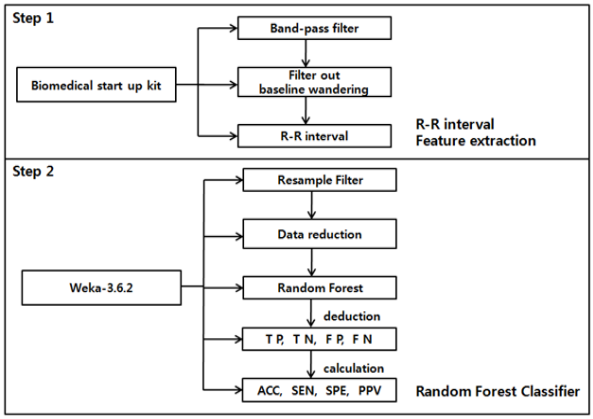


Figure 7. The process of the experiments

C. The results of the experiment

1) Accuracy comparison on the modification of the algorithm of Random Forest classifier between pre-results and now

In this study, an algorithm of RF (Random Forest) was modified to chase preferentially selected signals in order to improve its accuracy of results. Fig. 8 indicates differences between before its modification and after.

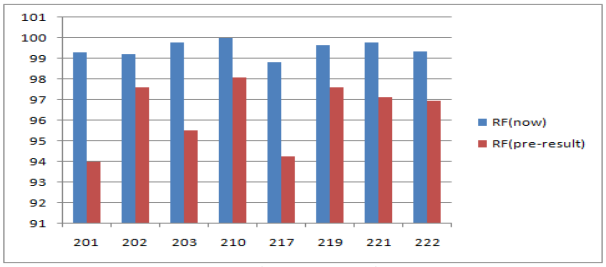


Figure 8. Accuracy comparison of the results of Random Forest algorithm modification between before and after

Data were selected based on consequences of K&L's experiment [27]. The sequence of the results stated at the Fig 8 is from RF (now) to RF (pre-result): RF (now) is results of the experiment after the modification, RF (pre-result) is results before the modification. From the Fig. 8, the algorithm after the modification shows more accurate than its counterpart before the modification.

2) Accuracy comparision among established researches, SVM, MLP and Random Forest

Table 1 indicates extracted results of Sensitivity, Specificity and Accuracy on R-R interval experiments which were conducted by established researches and Random Forest classifier.

 ISBN: 978-1-61208-203-5

TABLE 1. MEASUREMENT OF SENSITIVITY (SEN), SPECIFICITY (SPE) AND ACCURACY (ACC)

Arrhythmia Data		Sensitivity(%)		Specificity(%)		Accuracy(%)	
		K&L	*RF*	*K&L*	*RF*	*K&L*	*RF*
Record_ no	*201*	96.44	99.71	39.52	88.23	65.46	99.28
	202	80.79	99.62	94.64	72.73	88.72	99.2
	203	81.36	99.89	21.92	71.43	63.37	99.77
	210	96.89	100	0	0	94.73	100
	217	72.78	99.51	94.22	27.27	90.86	98.81
	219	96.86	99.81	64.20	89.66	91.39	99.63
	221	92.25	99.86	50.66	94.87	65.24	99.76
	222	92.25	100	50.66	0	65.24	100
Average		90.226	99.75	61	63.1	82.14	99.55

TABLE 3. ACCURACY ANALYSIS AMONG RF, MLP AND SVM

Record_no	Accuracy(%)		
	RF	*MLP*	*SVM*
201	99.28	83.69	81.27
202	99.2	96.42	96.23
203	99.77	80.42	77
210	100	96.38	95.69
217	98.81	75.95	68.45
219	99.63	92.89	92.89
221	99.76	86.36	82.31
222	100	77.34	74.87
Average	99.55	86.18	83.58

TABLE 2. RESULTS ON FEATURE DATA OF R-R INTERVAL EXPERIMENTED BY RANDOM FOREST CLASSIFIER

Record	TP	TN	FP	FN	PPV	SEN	SPE	ACC
201	1748	60	8	5	99.54	99.71	88.23	**99.28**
202	2081	24	9	8	99.56	99.62	72.73	**99.2**
203	2972	10	4	3	99.86	99.89	71.43	**99.77**
210	2352	25	0	0	100	100	0	**100**
217	2246	6	16	11	99.29	99.51	27.27	**98.81**
219	1600	26	3	3	99.81	99.81	89.66	**99.63**
221	2066	37	2	3	99.9	99.86	94.87	**99.76**
222	2567	0	0	0	100	100	0	**100**
Average	2204	23.5	5.25	4.125	99.745	99.8	55.52	**99.55**

Table 1 indicates the results of K&L, RF: K&L is Tateno's experiment [27], RF is this study's experiment. When it comes to comparison among them, it is stated below:

While the result of the 210 sector is the best in K&L's experiment, RF shows better performance as high as 5.27% compared to K&L.

Table 2 represents results that 8 number of feature data were experimented by Random Forest (RF).

After values of TP, TN, FP and FN had been previously extracted, PPV (Positive Prediction Value percentage), SEN (Sensitivity percentage), SPE (Specificity percentage) and ACC (Accuracy percentage) were measured, and then, the Average was calculated. Including those data, all of the other data showed over 90% of accuracy rates as well.

Thus, it could be regarede that Random Forest classifier extracted efficient Accuracy in the results of total data.

In order to analyze the accuracy of RF, Table 3 shows Accuracy rates among SVM, MLP, and RF. The sequence of the table is in order; RF -> MLP -> SVM. Through Table 3, it could be obviously comprehended that the accuracy of RF reaches approximately 100% compared to others.

3) Results cpmparison of HRV experiments between this study and established researches

Asl [4], which is the comparison of HRV (Heart Rate Variability) experiments on HRV, was used SVM (Support Vector Machine) and MLP (Multilayer Perceptron), and conducted by two ways; one was the case that GDA (Generalized Discriminant Analysis) which is 'Dimension reduce' method was adapted into the experiment and another was the 'GDA' was not adapted (no GDA) in common.

Thus, in this study, experiments were undertaken by two ways called 'All data' and 'Shorten' methods: total 25 number of properties were used in 'All data' method and the only 13 number of properties were used in 'Shorten' method. And Random Forest, MLP and SVM were commonly selected as algorithms of the experiments. Results of the experiments were then differently compared by cases: the results of 'All data' was compared with 'no GDA' [4] and its counterpart of 'Shorten' was compared with 'GDA'.

From the results of the experiment, 'GDA' shows better performance on 'Accuracy (ACC)' than 'no GDA' on HRV and SVM at 0.27% and 0.67% respectively. From the case of this study's consequence, 'Shorten' method indicated higher 'Accuracy' on MLP and SVM at 1.05% and 3.12% respectively.

TABLE 4. Results comparison of the experiment on HRV

Method		SEN (%)	SPE (%)	ACC (%)	Average
MLP	No GDA	90.64	98.51	98.22	95.79
	GDA	92.63	98.98	98.49	96.7
	Shorten	100	90.9	98.96	96.62
	All data	100	83.33	97.91	93.746
SVM	No GDA	92.57	98.88	98.49	96.646
	GDA	95.77	99.4	99.16	98.11
	Shorten	100	66.67	97.91	88.2
	All data	100	83.33	94.79	92.7
RF	Shorten	100	90.9	98.96	96.62
	All data	100	90.9	98.96	96.62

In terms of the comparison between 'GDA' and 'Shorten' method, although its 'Accuracy' was high at 0.47% when 'Shorten' method had adapted MLP, its 'Accuracy' was low at 1.25% when SVM was used. Finally, when RF (Random Forest) was adapted into the experiment, their 'Accuracy' (between the cases of MLP and SVM)

 ISBN: 978-1-61208-203-5

were at 98.96% at the same time. Why this study's results did not show remarkably higher performance than Asl [4] could be assumed that there was the lack of efficiency in the experiments of this study compared to Asl [4]'s research on 'pre-processing' as well as 'Dimension reduction' of the data.

V. CONCLUSION AND DIRECTION OF CONTINUOUS STUDY

In this study, the Accuracy rates among SVM, MLP and Random Forest classifiers were adapted into the comparative analysis of their performances using MIT-BIH Arrhythmia Database. Biomedical Startup Kit used the data extraction of R-R interval at pre-processing phases and Random Forest classifier provided by WEKA was used with its algorithm modified. In order to emphasize differences between the two groups, the algorithm of Random Forest classifier was modified by below three steps:

- The algorithm was changed to select high-frequent signals previously instead of random selection
- The algorithm was corrected to detect arrhythmia signals in the best-fitted segregation and classify them.
- Best-First decision tree was applied instead of Decision Tree.

As a result, the accuracy of Random Forest classifier could be remarkably maximized, and classifier did not show only higher performance than SVM and MLP classifier, but could also minimize out-of-bags. And, it was proved that the modified algorithm presented higher accuracy rates compared with the results of K. Tateno's researches in the aspect of the accuracy.

Consequently, despite of that remarkably high results were gained on the improvement of 10% accuracy rate, there were lower results at pre-processing phase than B. M. Asl's research process in terms of the next areas; exceeding limitation on Dimension reduction and used Band-pass Filter as well as efficient section separation of R-R interval. Therefore, after this study, it should be researched to select filter which can efficiently erase baseline-wandering in pre-processing phase and investigate methods that can accurately extract R-R interval.

ACKNOWLEDGMENT

This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2009.

REFERENCES

[1] K. S. Park, B. H. Cho, D. H. Lee, S. H. Song, J. S. Lee, Y. J. Chee, I. Y. Kim, and S. I. Kim, "Hierarchical Classification of ECG Beat Using Higher Order Statistics and Hermite Model," *K Kor Soc Med Informatics*, vol. 15, pp. 117-131, 2009.

[2] Korean Geart Rhythm Society. [Online]. Available (April 13, 2012): http://www.k-hrs.org/

[3] M. G. Tsipouras, D. I. Fotiadis , and D. Sideris, "An arrhythmia Classification system based on the RR-interval signal," *Artificial Intelligence in Madicine*, vol. 33, pp. 237-250, 2005.

[4] B. M. Asl, S. K. Setarehdan, and M. Mohebbi, "Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal," *Artificial Intelligence in Medicine*, vol. 44, pp. 51-64, 2008.

[5] P. D. Chazal, M. O' Dwyer, and R. B. Reilly, "Automatic Classification of Heartbeats Using ECG Morphology and Heartbeat Interval Features," *IEEE Transactions on Biomedical Engineering*, vol. 51, no.7, pp. 1196-1206, 2004.

[6] Michael and L. Hilton, "Wavelet and Wavelet Packet Compression of Electrocardiograms," *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 5, pp. 394-402, 1997.

[7] J. P. Martinez, R. Almeida, and S. Olmos, "A Wavelet-Based ECG Delineator: Evaluation on Standard Databases," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 4, pp. 570-581, 2004.

[8] I. Markovsky, A. Amann, and S. V. Huffel, "Application of Filtering Methods for Removal of Resuscitation Artifacts from Human ECG Signals," *30th Annual International IEEE EMBS Conference*, pp. 13-16, 2008.

[9] S. A. Taouli and F. B. Reguig, "Detection of QRS Complexes in ECG Signals Based on Empirical Mode Decomposition," *Global Journal of Computer Science and Technology*, vol. 11, pp. 11-17, 2011.

[10] H. Gholam-Hosseini, H. Nazeran, and K. J. Reynoids, "ECG Noise Cancellation Using Digital Filters," *2nd International Conference on Bioelectromagnetism*, pp. 151-152, 1998.

[11] K.C. Chua, V. Chandran, R.U. Acharya, L.C. Min, "Cardiac Health Diagnosis Using Higher Order Spectra and Support Vector Machine," *Open Med Inform J 3*, pp. 1–8, 2009.

[12] M. Bsoul et al., "Real-time sleep quality assessment using single-lead ECG and multi-stage SVM classifier," *in the International Conference of IEEE Engineering in Medicine and Biology Society*, pp. 1178–1181, 2010.

[13] H. Zhang and L. Q. Zhang, "ECG analysis based on PCA and Support Vector Machines," *Networks and Brain, IEEE*, pp. 743-747, 2005.

[14] O. T. Inan and G. T. A. Kovacs, "Robust Neural-Network-Based Classification of Premature Ventricular Contractions Using Wavelet Transform and Timing Interval Features," *IEEE transctions on biomedical engineering*, vol. 53, no. 12, 2006.

[15] F. Yaghouby, A. Ayatollahi, R. Soleimani, "Classification of Cardiac Abnormalities Using Reduced Features of Heart Rate Variability Signal," *World Applied Sciences Journal*, vol. 6, no. 11, pp. 1547–1554, 2009.

[16] Y. Özbay, R. Ceylan and B. Karlik, "A fuzzy clustering neural network architecture for classification of ECG arrhythmias," *Comput. Biol. Med.*, vol. 36, pp. 376–388, 2006.

[17] J.R. Quinlan, "C4.5: Programs for machine learning," *Morgan Kaufmann Publishers, San Francisco CA*, 1993.

[18] T.P. Exarchos, M.G. Tsipouras, C.P. Exarchos, C. Papaloukas, D.I. Fotiadis and L.K. Michalis, "A methodology for the automated creation of fuzzy expert systems for ischaemic and arrhythmic beat classification based on a set of rules obtained by a decision tree," *Artif. Intell. Med.*, vol. 40, no. 3, pp. 187–200, 2007.

[19] V. Mahesh, A. Kandaswamy, C. Vimal and B. Sathish, "Cardiac disease classification using heart rate signals," *Int. J. Electronic Healthcare*, vol. 5, no. 3, 2010.

[20] PhysioBank. [Online]. Available (April 13, 2012): http://physionet.mit.edu/physiobank/database/mitdb/

 ISBN: 978-1-61208-203-5

[21] NI Biomedical Startup Kit 3.0. [Online]. Available (April 13, 2012): http://decibel.ni.com/content/docs/DOC-12646

[22] G. D. Clifford, F. Azuaje, and P. E. McSharry, "Advanced Methods and Tools for ECG Data Analysis," *Artech House*, pp. 101-102, 2006.

[23] All About Circuits URL. [Online]. Aailable (April 13, 2012): http://www.allaboutcircuits.com/

[24] L. Breiman, "Machine Learing," *Kluwer Academic Publishers in Netherlands*, 2001.

[25] P. N. Tan, M. Steinbach and V. Kumar, "Introduction to Data Mining," *1-st Addison-Weseley*, 2006.

[26] H. Shi, "Best-first Decision Tree Learning," *The University of Waikato in NewZealand*, pp. 3-5, 2007.

[27] K. Tateno and L. Glass, "A Method for Decision of Atrial Fibillation Using RR-Intervlas," *Computers in Cardiology(IEEE)*, vol. 27, pp. 391-394, 2000.

 ISBN: 978-1-61208-203-5

Scalable Democratic Routing in Wireless Mesh Networks

Ronit Nossenson
Faculty of Computer Science
Jerusalem College of Technology (JCT)
Jerusalem, Israel
nossenso@jct.ac.il

***Abstract* — We propose a new method for scalable routing in large wireless mesh network: the democratic routing scheme. In this new schema, the nodes are divided into components according to their connectivity classes. As oppose to hierarchical routing, here, all nodes in the component are equal. The routing decisions are performed according to nodes connectivity structure together with a proper routing performance metric. Every node holds a view including its neighbor set and a dynamic connectivity model of the network. The node uses the view to understand which of its topology changes should be announced and to identify the set of nodes that should get this specific update. In this way, the routing overhead is significantly reduced, and, yet, the necessary routing information is available.**

***Index Terms* — Wireless mesh network; scalable routing algorithms; connectivity models.**

I. INTRODUCTION

Various wireless networks have evolved into the next generation to provide better services. A key technology, Wireless Mesh Networks (WMN), has emerged recently [1]. A WMN is dynamically self-organized and self-configured, with the (possibly mobile) nodes in the network automatically establishing and maintaining mesh connectivity among themselves. In such systems, the implied routing protocol directly affects scalability, efficiency, and reliability.

When the network is large, hierarchical routing protocols tend to achieve better performance [2-10]. Hierarchical routing protocols build a hierarchy of nodes, typically through clustering techniques. These protocols divide the nodes in the network into backbone nodes and regular nodes arranged in clusters. Every cluster uses a *cluster head* node that is a part of the backbone. The cluster head node acts as a local coordinator of transmissions within the cluster and is responsible for keeping and updating routing information beyond the cluster. However, explicit clustering schemes have several drawbacks [3]:

- Clustering usually requires that the cluster heads will be more powerful (battery life, transmission range and capacity) than the regular nodes in the clusters. Unless being intentionally designed so, the cluster head may become a bottleneck.
- The complexity of maintaining the hierarchy compromises the performance of the routing protocol. There is a significant overhead in the maintenance of the cluster (e.g., electing the cluster head and maintaining the cluster's members);
- The centralization and load of the cluster-heads routes;
- The routing path may be longer than a direct path;
- Clustered protocols are sensitive to failures of the cluster heads;
- The fragmentation of the network nodes into clusters may result in a large number of clusters.
- Change of cluster heads results in routing changes and hence generates routing overhead.

In this research, we propose a completely different method for scalable routing in large WMN, the *democratic routing scheme*. In this new schema every node holds a *view*. The view includes the node neighbor set and a *dynamic connectivity model* of the network. Using its view, the node can keep updated information on the network and can perform educated routing decisions. The dynamic connectivity model represents the network topology in a compact structure (O(n), where n is the number of nodes in the network) and can be updated in an efficient manner to capture the dynamic changes in the network topology, as nodes move from one place to another while establishing and releasing links. On one hand, this model should be small to avoid unnecessary updates (meaning, high overhead); on the other hand it should be detailed enough to allow good routing decisions. The dynamic connectivity model is updated whenever an *essential* change in the network accrues. The node uses the view to understand which of its topology changes (for example, link establishment or link release) should be announced and what is the exact set of nodes that should get the specific update. In this way the routing algorithm overhead is significantly reduced and, yet, the necessary routing information is available.

We choose to describe our new concept using an extension of the cactus-tree model of Dinitz et al. [11] and the two-level cactus-tree model [12] to enjoy their elegancy and simplicity. However, these models support incremental maintenance only and they do not support node-deletion and edge-deletion (link release). Thus, we extend the models to support node-deletion and edge-deletion for low connectivity levels.

This paper is organized as follows. In the next section, related works are listed. In section III, basic definitions are described. Section IV presents the connectivity model dynamics. Section V describes the new democratic routing scheme. Finally, conclusions and further research topics are given in Section VI.

 ISBN: 978-1-61208-203-5

II. Related Work

Typically, when wireless network size increase flat routing schemes become infeasible because of link and processing overhead. One way to solve this problem and to produce scalable solutions is hierarchical routing. The common way of building hierarchy is to group nodes geographically close to each other into explicit clusters. In explicit clustering schemes, each cluster has a leading node (cluster head) to communicate to other nodes on behalf of the cluster. An alternate way is to have implicit hierarchy. In this way, each node has a local scope. Different routing strategies are used inside and outside the scope.

Cluster head-Gateway Switch Routing (CGSR) [7] is typical of explicit cluster-based hierarchical routing. A stable clustering algorithm, Least Cluster head Change (LCC), is used to partition the whole network into clusters, and a cluster head is elected in each cluster. A mobile node that belongs to two or more clusters is a gateway connecting the clusters. Packets are routed through paths having a format of "Cluster head–Gateway Cluster head–Gateway…" between any source and destination pairs.

Additional well known explicit clustering routing protocol is the Hierarchical State Routing (HSR) [5]. It is a multilevel clustering-based Link State routing protocol. It maintains a logical hierarchical topology by using the clustering scheme recursively. Nodes at the same logical level are grouped into clusters. The elected cluster heads at the lower level become members of the next higher level. These new members in turn organize themselves in clusters, and so on. The cluster head summarizes link state information within its cluster and propagates it to the neighbor cluster heads (via the gateways).

The basic idea in the Zone Routing Protocol (ZRP) [6] is that each node has a predefined zone centered at itself in terms of number of hops (implicit cluster). For nodes within the zone, it uses proactive routing protocols to maintain routing information. For those nodes outside of its zone, it does not maintain routing information in a permanent base. Instead, on-demand routing strategy is adopted when inter-zone connections are required.

A recent study [3] considers a routing technique which can implicitly cause nodes that are in the "center" of dense areas to act as cluster heads ("natural clustering"). Using the Metrical Routing Algorithm (MRA) [4], it maintains a dynamic set of coordinates to every node. Thus, if the coordinates of the destination are known, the MRA sends a message to this destination through the shortest path based on the estimated metrical distances.

We propose a new democratic routing scheme which can be classified as an implicit clustering. In this scheme, the nodes are logically divided into their connectivity classes (components). The components do not have component heads and the traffic is handled in a complete democratic manner. That is, in our scheme all nodes are equal. The routing decisions are performed according to nodes connectivity structure together with a proper routing performance metric (for example, minimum hops metric). Unlike other implicit clustering schemes, in our solution, nodes do not have geographic location information on the other nodes (coordination). In addition, we use the same routing protocols for routing inside and outside the connectivity components.

III. Definitions

Let G=(V,E) be a weighted undirected connected multi-graph without loops induced from a specific network topology, where every vertex represents a node in the network and every edge between two vertices represents a wireless link between a pair of corresponding nodes that are in communication range.

A minimal edge-cut C of G is an edge set whose removal disconnects G and removal of any proper part of C does not disconnect G. If |C|= k then C is called a k-cut. If C={e} (that is |C|=1) then the edge e is called a *bridge*. Two vertices {u,v} are called k-edge-connected if no k'-cut, k' < k, separates u from v. It is well known that the property ``there exist k edge-disjoint paths between u and v in G" defines the same relation as k-edge-connectivity. The equivalence classes of this relation are called the k-edge-connected classes (*k-classes* for short). The partition of V into the (k+1)-classes is a refinement of the partition of V into k-classes. Thus, the connectivity classes have a hierarchical structure.

For a k-connected graph, its connectivity model represents both its (k+1)-classes and its k-cuts. For example, the well known bridge-tree model of a 1-connected graph represents its 1-cuts (the so-called *bridges*) and its 2-classes [16]. Similar, the cycle-tree connectivity model of a 2-connected graph represents its 2-cuts and its 3-classes [13] [14]. These connectivity models are, in fact, special cases of a more general connectivity model called the cactus-tree model [11]. The cactus-tree model [11] of a k-connected graph represents both its (k+1)-classes and its k-cuts.

For the simplicity of this short presentation, we assume that the network is not highly connected and use the bridge-tree model of [16] together with the cycle-tree model of [14] [15] for each 2-class in the graph. This joined two-level connectivity model is, in fact, a special case of the two-level cactus-tree model of [12]. Using the general model of [12], our result can be easily adjusted to highly connected networks.

By definition, the size of this connectivity model is O(n), where n is the number of nodes in the network. If n is very large, it is possible to divide the network according to geographic location and to define proper gateways nodes to connect the networks parts.

Let S be a sub-set of V. The induced graph G(S) consists of the vertices S and all edges in E connecting vertices in S. For a 3-class S, the associated 3-componnet graph is the induced graph G(S) together with virtual edges. The virtual edges represents cycles in the cycle-tree model that are attached to distinct vertices of S. The 3-commponent mimics the connectivity structure of S in a localized fashion [13].

 ISBN: 978-1-61208-203-5

IV. CONNECTIVITY MODEL DYNAMICS

Each node in the network holds a *view*. The view of node v includes:

- The global connectivity model of the network (the bridge-tree and the cycle-tree of each 2-class),
- Its local connectivity model: the cactus-tree model of v's 3-componnent graph.
- The set of v's neighbors

In this section, we describe the dynamics of the connectivity model. The *vertex insert* and *edge insert* procedures are given in [12]. Here, we provide an intuitive explanation, and present two examples for the model completeness. For formal description of these procedures, see [12] [14]. We add new procedures to support *vertex delete* and *edge delete*. Note that the new transformations described here are designed for the simple case of low connectivity only.

Regarding a new vertex insertion, the vertex is inserted as a singleton. Meaning, until edge insertion it act as an isolated node. The corresponding connectivity models are trivial.

Regarding a new edge e=(u,v) insertion (new link establishment), there are four cases according to the relation between the vertices u and v. The four cases are: u and v belong to distinct 1-classes, u and v belong to the same 1-class but to distinct 2-classes, u and v belong to the same 2-class but to distinct 3-classes, u and v belong to the same 3-class.

The change in the connectivity models due to new link establishment between two nodes belonging to separated networks (1-classes) is a simple connection of the two representing models by adding a new bridge associated with the new link e. To connect nodes with higher connectivity a simple *squeeze* operation is performed on the path-of-edges-and-cycles, in which the set of all nodes along the path are replaced by a single node [12] [14].

The following new procedure defines the vertex deletion operation.

```
Void VertexDelete(v)
Begin
1: for every edge e attached to v do {
2:   remove edge (e); }
3: release vertex v;
End
```

This procedure functionality includes releasing of all the edges attached to this vertex and then releasing the vertex resources (memory etc.).

Theorem 1

The procedure VertexDelete(v) correctly updates the connectivity models.

Regarding an edge deletion operation (see EdgeDelete procedure below), we have three cases. Assume that the edge e=(u,v) needs to be deleted. First, the vertices u and v can belong to the same 1-class but to distinct 2-classes. Second, the vertices u and v can belong to the same 2-class but to distinct 3-classes. In the third case, u and v belong to the same 3-class.

Consider the first case (u and v belong to the same 1-class but to distinct 2-classes).

Lemma 2

Assume that e=(u,v), u and v belong to the same 1-class but to distinct 2-classes. Then, e is a bride in the graph and has direct representation in the bridge-tree model.

Proof

By bridge definition.

In this case (steps 1-3 in the EdgeDelete procedure below), the global connectivity model is changed by removing the edge which represents e in the model. The local connectivity models of u and v are not affected since u and v do not belong to the same 3-class (component).

Consider the second case (u and v belong to the same 2-class, but to distinct 3-classes).

Lemma 3

Assume that e=(u,v), u and v belong to the same 2-class but to distinct 3-classes. Then, e has direct representation in the cycle-tree model of the 2-class.

Proof

Since u and v belong to the same 2-class but to distinct 3-class there are exactly two edge-paths between u and v. One path consists of {e} and let us denotes the second one by P. Removing e together with any edge from P result in a minimal cut that separates u and v and must be represented in the cycle-tree. That is, the cycle L= {e}U{P} is in the cycle-tree.

In this case (steps 4-6 in the EdgeDelete procedure below), the global connectivity model is changed by transforming the cycle which include the edge which represents e into a path of bridges by removing this edge. The local connectivity models of u and v are not affected since u and v do not belong to the same 3-class and to the same 3-componnent. In addition, by definition, the cycle L is not represented by a virtual edge in the 3-componnets of u and v.

Consider the third case (u and v belong to the same 3-class, denoted by S). Removing an edge in this case might result a fragmentation of the 3-class S. Formally, now we have two cases: u and v can belong to the same 4-class or they can belong to distinct 4-classes. If u and v belong to the same 4-class (steps 18-23 in the EdgeDelete procedure below) then the global connectivity model will not change (the class S is still a 3-class) as a result of removing e. Only the local cactus-tree might change as follows. The node that represents S is replaced with a cactus-tree (*implanting* the model instead of the node). This is done via 3-componnent discovery operation in addition to virtual edges that result from the cycle-tree.

In the second case (steps 7-17 in the EdgeDelete procedure below), the node that represents S in the global

connectivity model, should be replaced with a cycle-tree that will represent S as a 2-class (*implant* the model instead of the node). That is, if u and v belong to distinct 4-classes then the removal of e will change the 3-class S and turn it to a 2-class. The cycle-tree that represents S is generated from the cactus-tree model of the 3-componnent associated with S in the following manner. Let T be a path in the cactus-tree between the node that represent the 4-class of u and the node that represents the 4-class of v. Every 3-cut that is represented by an edge on this path is now a 2-cut. Thus, the cycle-tree of S is a sequence of cycles (each of size 2). All nodes in the cactus-tree before the path T stay in one 3-class, and all nodes after this path stay in one 3-class. Each node in the path T is now a 3-class. In addition, each node in these new 3-classes should construct a new local connectivity model: the cactus-tree of its new 3-class. This is done via 3-componnent discovery operation in addition to virtual edges that result from the cycle-tree.

```
Void EdgeDelete(u,v)
Begin
1: if (u and v belong to the same 1-class
     but to distinct 2-classes) do {
2:   Remove e from the bridge-tree;
3:   Update u and v global models;}
4: else if (u and v belong to the same
     2-class but to distinct 3-classes) do
{
5:   Remove e from the cycle-tree;
6:   Update u and v global models;       }
7: else if (u and v belong to the same
     3-class but to distinct 4-classes) do
{
8:   find the path between u's 4-class and
     v's 4-class in the cactus-tree;
9:   create the proper cycle-tree from the
     cactus-tree;
10:  implant the new cycle-tree in the
     global model;
11:  Update the global models of all nodes
     in the 3-class;
12:  for every new 3-class created do{
13:     discover the 3-componnet;
14:     add virtual edges according to the
          cycle-tree;
15:     calculate the cactus-tree;
16:     for every node in this 3-class do{
17:        Update the local model; } } }
18: else if (u and v belong to the same
     4-class) do {
19:  discover the 3-componnet;
20:  add virtual edges according to the
          cycle-tree;
21:  calculate the cactus-tree;
22:  for every node in this class do {
23:  Update the local model; } }
End
```

The example plotted in Figure 1 describes a delete edge operation, where e=(u,v) and the vertices u and v belong to the same 3-class (denoted by N_2 in the figure). The induced graph is presented in (a); the global connectivity-model of this network is presented in (b), it consists of all 1 and 2 – minimal cuts of the graph. In (c), the 3-componnent corresponding to the 3-class N_2 is presented together with the local model of the vertices in the 3-class N_2. In addition, we can see the transformation of this 3-class due to the removal of the edge e, with the proper transformation of its cactus-tree model (representing the minimal 3-cuts of 3-class N_2) into a cycle-tree (representing the minimal 2-cuts of 3-class N_2). Finally, (d) shows the updated global connectivity model.

Theorem 4

The procedure EdgeDelete(v) correctly updates the connectivity models.

V. The Democratic Routing Scheme

In this section, we describe the democratic routing scheme. As mentioned before, each node has its view. The node uses the view to decide which of its topology changes should be announced and the exact set of nodes that should get the specific update.

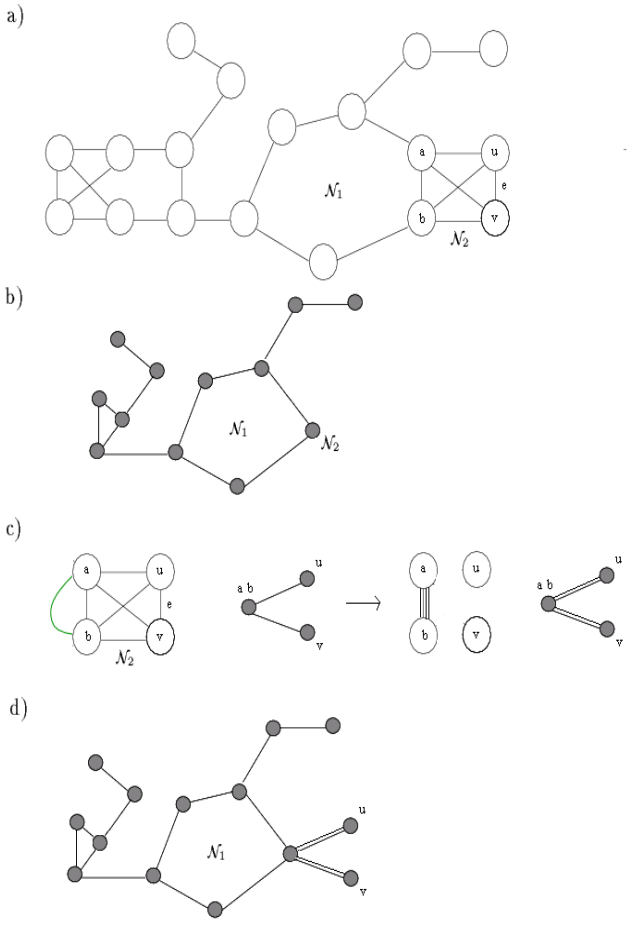


Figure 1. Deleting an edge between vertices belonging to the same 3-class but to distinct 4-classes

When a new node joins the network it performs an Insert vertex operation. As a result, its view is initiated. Next, the node discovers its neighbors. For each link establishment between the node and an adjusted node, the node received the neighbor's global and local connectivity model. Once an update is required on the global connectivity model, an 'update global model' message is broadcast to all nodes. This message includes the origin node identity, a sequence number generated in the origin node, and the update description. An important observation is that a node which receives two distinct 'update global model' messages can join these messages into one message including both updates and by that to reduce the routing overhead.

Once an update is required on the local connectivity model, an 'update local model' message is send to the nodes in the 3-componnent only. If there are no updates to the connectivity models, no updates messages are sent. In component discovery message, each node which belong to the specific 3-class responses with its neighbors set.

As a node move from one place to another its links changes. Some new links are added, some old links are been released. The node continues to update the necessary routing information and the truly important network topology changes are continually monitored.

Since the connectivity models can be decomposed according to the network topology, it is possible to divide the network to sub-networks and to define the models for each part. Using special nodes as gateways can solve the problem of connecting the sub-networks.

VI. CONCLUSION AND FUTURE WORK

In this on-going research, we proposed a new democratic routing scheme. In this scheme all nodes are equals and routing decisions are based on the nodes view of the network. To limit the routing algorithm overhead, the nodes use connectivity model to decide which node should be informed of a topology change if any.

Future work includes implementation of this scheme and evaluation of the algorithm performance comparing to others scheme for scalable routing in large wireless mesh networks.

REFERENCES

[1] Akyildiz, I., F., Wang, X., and Wang W., "Wireless mesh networks: a survey", Computer Networks 47 (2005) pp. 445–487, Jan 2005.

[2] Belding-Royer, E., M., "Multi-level hierarchies for scalable ad hoc routing", ACM/Kluwer Wireless Networks vol. 9 issue 5 pp. 461–478, 2003.

[3] Ben-Asher, Y., Feldman, S., Feldman M., and Gurfil, P., "Scalability Issues in Ad-Hoc Networks: Metrical Routing Versus Table-Driven Routing", Wireless Pers Commun (2010) 52:423–447.

[4] Ben-Asher, Y., Feldman, S., and Feldman, M. "Ad-hoc routing using virtual coordinates based on rooted trees", In IEEE SUTC, Taiwan, 2006.

[5] Chiang, C., and Gerla, M., "Routing and Multicast in Multihop, Mobile Wireless Networks," Proc. IEEE ICUPC '97, San Diego, CA, Oct. 1997.

[6] Haas, Z. J. and Pearlman, M. R., "The Performance of Query Control Schemes for the Zone Routing Protocol," ACM/IEEE Trans. Net., vol. 9, no. 4, Aug. 2001, pp. 427–38.

[7] Pei, G. et al., "A Wireless Hierarchical Routing Protocol with Group Mobility," Proc. IEEE WCNC '99, New Orleans, LA, Sept. 1999.

[8] Hagouel, J., "Issues in Routing for Large and Dynamic Networks", PhD thesis, Columbia Univ., May 1983

[9] Saha, A., K., Johnson, D., B., "Self-organizing hierarchical routing for scalable ad hoc networking", Technical Report, TR04-433, Department of Computer Science, Rice University

[10] Tsuchiya, P., F., "The landmark hierarchy: A new hierarchy for routing in very large networks", ACM SIGCOMM Comput. Commun. Review 18, 4, pp. 35–42, August 1988.

[11] Dinic, E., A., Karzanov A., V., and Lomonosov, M. V., "On the structure of the system of minimum edge cuts in a graph", Studies in Discrete Optimization, A. A. Fridman (Ed.), Nauka, Moscow, 1976, pp. 290-306 (in Russian).

[12] Dinitz, Ye., and Nutov, Z., "A 2-level cactus tree model for the minimum and minimum+1 edge cuts in a graph and its incremental maintenance", Proc. the 27th Symposium on Theory of Computing, 1995, pp. 509-518.

[13] Dinitz, Ye., "The 3-edge components and the structural description of all 3-edge cuts in a graph", Proc. 18th International Workshop on Graph-Theoretic Concepts in Computer Science (WG92), Lecture Notes in Computer Science, v.657, Springer-Verlag, 1993, pp. 145-157.

[14] Dinitz Ye., and Westbrook, J., "Maintaining the Classes of 4-Edge-Connectivity in a Graph On-Line", Algorithmica, Volume 20, Number 3, 1998, pp. 242-276.

[15] Galil, Z., and Italiano, G., F., "Maintaining the 3-edge-connected components of a graph on line", SIAM J. Computing 22(1), 1993, pp. 11-28.

[16] Westbrook, J., and Tarjan, R., E., "Maintaining bridge-connected and biconnected components on line", Algorithmica, 7, 1992, pp. 433-464.

 ISBN: 978-1-61208-203-5

Demand Aware Fair Resource Allocation in TDMA Wireless Networks

Xiaolong Huang and Soumya Das
QUALCOMM Inc., San Diego, California, USA
Email: {xhuang, soumyad}@qualcomm.com

***Abstract*—Meeting traffic demand and enforcing fairness are often times necessary but conflicting objectives for resource allocation in wireless networks. Due to the resource sharing nature of wireless networks, without a mechanism for enforcing fairness, simply assigning resources to meet traffic demand of some network flows can lead to resource starvation of other network flows. Balancing these two objectives is more complex multi-hop wireless networks, as the resource contention could be indirect. In this paper, an algorithm is introduced to allocate time slots in TDMA-based multi-hop wireless networks to achieve a designated a balance between meeting traffic demand and enforcing fairness. Numerical results show that the algorithm performs significantly better than other resource allocation algorithms. The introduced algorithm is well suited for distributed TDMA-based wireless networks, such as ECMA-368 based UWB networks.**

***Index Terms*—Congestion control, fairness, multi-hop wireless network, optimization, quality of service, resource allocation, TDMA.**

I. Introduction

Quality of Service (QoS) and fairness are both important yet often times mutually conflicting objectives for resource allocation and scheduling in wireless networks. Due to the resource sharing nature of a wireless environment, meeting the QoS of some flows without addressing the fairness issue may lead to resource starvation of other flows. The problem of balancing QoS and fairness becomes more complex when the network spans more than a single hop. This is evident in Fig. 1.

In Fig. 1, a flow contention graph is composed of three one hop flows $\{A, B, C\}$. Due to some underlying network topology, flow B contends with both flows A and C, while flows A and C do not contend with each other directly. Hence, a transmission of either flow A or flow C would block flow B. Two time slots are assumed available for the three flows to use. Each flow is assumed to require one time slot to meet its traffic demand. As we can see, the resource allocation strategy at the top leaves flow B no time resource to use, while the resource allocation strategy at the bottom can serve all traffic demands.

The problem of maximizing the time slot allocation efficiency in TDMA wireless networks by exploiting the spatial reuse is NP-complete [8]. Several algorithms [8], [10] have been introduced to probabilistically achieve the maximum resource allocation efficiency without considering QoS and fairness.

QoS and fairness of resource allocation in wireless networks have been studied in separate contexts extensively. Various

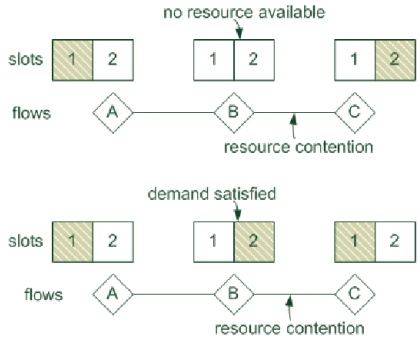


Fig. 1. Resource Allocation Problem in Multi-hop Wireless Networks

fairness measures have been introduced. Some solutions are designed to achieve specific objectives, such as proportional fairness [1] and max-min fairness [2]. Algorithms to achieve those objectives are usually complex. Other resource allocation algorithms, such as DRAND [4], provide a level of fairness and spatial reuse in multi-hop ad hoc networks without specific objective functions but involve much less computational complexity. These algorithms enforce fairness in the absence of QoS requirement, hence can be applied when every flow impose infinite traffic demand. On the other end, various resource allocation algorithms are introduced to solely meet the QoS requirements. Schemes, such as [5], allocate time slots on a flow by flow basis to meet their traffic demands and can easily lead to unfair congestion situations.

Several algorithms [9], [11] have been introduced to address the tradeoff between QoS and fairness by dynamically allocating time slots based on traffic loading and flow contention. In [3], a gradient method based resource allocation scheme is introduced to gradually regulate the data rate of end-to-end flows so that a utility function can be maximized across the network under the underlying flow contention constraint. Such schemes require adjusting allocated resources in a highly dynamic manner. However, demand assigned TDMA-based wireless networks, such as WiMedia networks [6], [7], expect such resource assignment to be static over a period of time. Hence, the aforementioned schemes are not suitable for such a deployment.

In this paper, a Demand Aware Fair resource allocation algorithm (DAF) is proposed to allocate time slots in TDMA-based multi-hop wireless networks. The DAF algorithm considers the requirement of QoS and fairness jointly, where QoS is

 ISBN: 978-1-61208-203-5

measured by the amount of traffic demand that is being served. The number of assigned time slots remains static during flow holding (active) times. To meet the traffic demand and to preserve fairness among multiple flows, DAF is designed to achieve the following objectives in TDMA-based multi-hop wireless networks:

- A flow is guaranteed a minimal number of time slots, called fair share when it has infinite traffic demand. This imposes a basic standard of fairness.
- The traffic demand of a flow is met when it is lower than the fair share of the flow.
- Achieve a prescribed balance between serving the traffic demand and reducing congestion.

We show that the proposed DAF algorithm meets the traffic demand and enforces the predefined fairness. DAF is well suited for distributed TDMA-based wireless networks, such as a WiMedia network, in which nodes advertise their available time slots to their 2-hop neighbors and then exchange messages to reserve time slots for new flows.

We introduce several important concepts in Section II. In Section III, the DAF algorithm and its objective is described. Numerical results are given in Section IV. Section V concludes this paper.

II. System Model

In this section, we present and develop several concepts for modeling the resource allocation problem in a TDMA-based multi-hop wireless network.

A. Maximal Common Slot Set

In this paper, a one-hop flow is simply referred to as a flow. A flow is always considered bidirectional so to take into account both data and acknowledgement transmissions.

In a TDMA-based wireless network, a flow f can only have transmissions in a time slot s during which its sender and receiver are not participating in transmissions for other flows. The time slot s is then said to be available to flow f.

Definition 1: A set of time slots S is said to be *commonly available* to a group of flows F, if $\forall s \in S$ and $\forall f \in F$, s is available to f. S is said be a *common slot set* of F. F is said to be a *common flow group* of S.

A time slots set S, a flows group F, and their relations can be modeled as a bipartite graph $G = (S + F, E)$. An edge exists between $s \in S$ and $f \in F$ if and only if s is available to f. A common slot set S and its common flow group F forms a complete bipartite graph.

Definition 2: A group of flows F is said to be the *maximal common flow group* of its common slot set S, if and only if S does not have another common flow group $\widetilde{F}$ so that $F \subset \widetilde{F}$.

Definition 3: A set of time slots S that has a maximal common flow group F is said to be a *maximal common slot set* (MCSS), if and only if 1) No other time slot set $\widetilde{S}$ has F as its maximal common flow group, and 2) No other time slot set $\widetilde{S}$, $\widetilde{S} \subset S$ has a maximal common flow group $\widetilde{F}$ such that $\widetilde{F} \supset F$.

The first requirement means that the complete bipartite graph formed by S and F includes all time slots that are available to F. The second requirement means that a maximal common slot set does not contain any common slot set that serves more flows than it does.

The significance of the maximal common slot set is that a TDMA frame can be partitioned into disjoint maximal common slot sets, where each MCSS set has a designated maximal flow group it can serve.

A simple way to obtain MCSSs is to start with individual time slots and their maximal common flow groups, and then group those time slots that have identical maximal common flow groups. The process stops when no two time slot sets have identical maximal flow groups. The maximum computational complexity is of $O(n^2)$, where n is the number of time slots. In this paper, we do not study the algorithm of obtaining maximal common slot sets.

B. Maximal Common Slot Set based Flow Contention Graph

In wireless networks, two transmissions may cause strong interference between each other if they overlap in time, frequency and space. In this paper, only single carrier TDMA-based multi-hop wireless networks are treated. Hence, only time and space domains can be explored, which manifests as TDMA operations and spatial reuse, respectively. To precisely capture the exploration of time and space, we introduce the concept of Maximal Common Slot Set based Flow Contention Graph (MCSS-FCG).

Definition 4: Two flows are said to be *contending flows* for each other if their simultaneous transmissions cause strong interference to each other and subsequently result in transmission failures. Under a protocol model, two flows are said to be contending for each other if the source or the destination of one flow is within the nominal communication range of that of the other flow.

A Flow Contention Graph (FCG) captures all flow contention information. The mapping from a nodal graph to a flow contention graph is well known [12] and illustrated in Fig. 2. In Fig. 2, all five flows are considered bidirectional. Each flow in the nodal graph is converted into a vertex in the flow contention graph. An edge exists between two vertices in a flow contention graph if and only if the corresponding two flows contend with each other.

Definition 5: A *maximal clique* is a set of vertices that induces a complete graph, and is not a sub-graph of any other complete graph. A *degree* of a maximal clique is defined as the number of vertices in that clique.

A flow contention graph can be decomposed into a set of maximal cliques as shown in Fig. 3. In Fig.3, the flow contention graph is composed of three maximal cliques, $\{A, B, C\}$, $\{B, C, D\}$ and $\{D, E\}$. Each clique is a complete graph and is not a sub-graph of any other complete graph in the flow contention graph.

Definition 6: A *Maximal Common Slot Set based Flow Contention Graph* (MCSS-FCG) is just a flow contention graph with respect to a maximal common slot set. All time

 ISBN: 978-1-61208-203-5

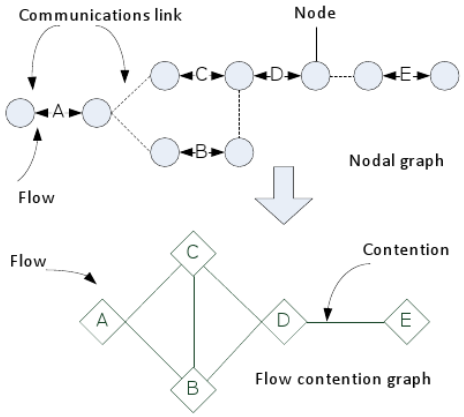


Fig. 2. Nodal Graph and its Flow Contention Graph

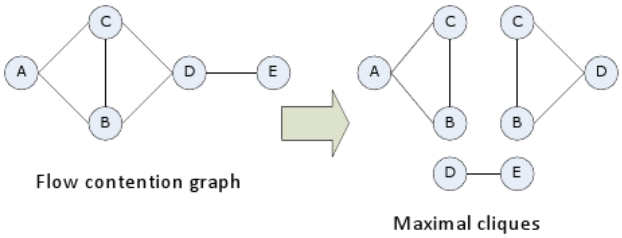


Fig. 3. Maximal Clique Decomposition of a Flow Contention Graph

slots in a maximal common slot set are available to all flows in the MCSS-FCG.

A MCSS-FCG is only meaningful with respect to its maximal common slot set. MCSS-FCGs associated with different maximal common slot sets are distinct. Flows that are assigned time slots in different maximal common slot sets do not contend with each other. We note that, when all time slots are available to all flows in the network, there exists a unique MCSS-FCG in the network, which is the overall flow contention graph itself.

An example of maximal common slot sets of their MCSS-FCG is illustrated in Fig. 4. There are 7 flows, $\{A, B, ..., G\}$, in the overall flow contention graph. Each flow has some particular time slots available in the frame for its use. Three maximal common slots are assumed to be identified. Each maximal common slot has a flow contention graph that is formed by its common flow group. A flow may reside in multiple MCSS-FCGs. For example, all three maximal common slot sets are available for flow A. Hence, flow A resides in all three MCSS-FCGs.

C. Fair Share in a MCSS-FCG

Definition 7: A *fair share* is defined as a number of time slots that shall be assigned to a flow when every flow in the network imposes infinite traffic demand that can saturate the network.

Assigning fair shares to flows when flows contend among one another in a traffic overloaded network enforces a standard of fairness. Since flows that are assigned with time slots in

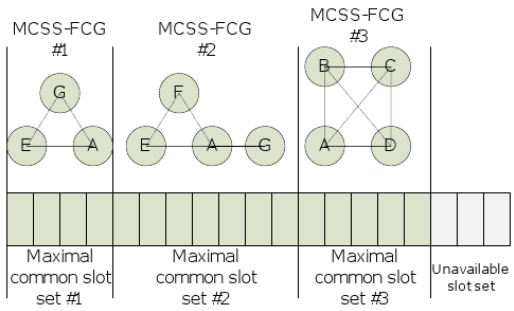


Fig. 4. Maximal Common Slot Set and its MCSS-FCG

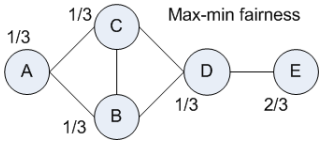


Fig. 5. Max-min Fair Share Assignment

different maximal common slot sets do not contend with each other, the fair share is only meaningful with respect to a specific MCSS-FCG. When a flow resides in multiple MCSS-FCGs, this flow has a fair share value in each MCSS-FCG.

Fair shares can be assigned based on a specific objective or assigned arbitrarily based on service agreement. To give an intuitive idea of fairness, the normalized max-min fair share assignment in a perfect graph studied in [2] is shown in Fig. 5. It is feasible to allocate and schedule $\frac{1}{3}$ time slots to Flows A, B, C, and D and allocate $\frac{1}{3}$ time slots to Flow E. The fairness nature of this allocation is that no flow can increase the assigned amount of time slots without reducing the time slots of other flows that are already assigned with less or equal amount of time slots.

In our resource allocation algorithm study, a simple rule of assigning fair shares is used. The fair share for a flow in a MCSS-FCG that has L time slots in the associated maximal common slot set is set to $\frac{L}{d}$, where d is the highest degree of all maximal cliques the flow resides in.

III. Resource Allocation Algorithm

At any point in time, a TDMA-based multi-hop wireless network may be serving existing flows while a new set of flows may be initiated in the network. The traffic demand of these new flows can be expressed in bits per frame. To negotiate the traffic demand to be served, a signaling process can compute and then reserve time slots for these flows on selected routes. Our DAF resource allocation algorithm used by the signaling process strives to achieve the following objectives:

- The portion of traffic demand within the fair share of a flow should be fully met.
- Minimize the cost associated with inadequate serving of traffic demand and the cost associated with allocating time slots above fair shares.

DAF is executed over a set of MCSS-FCGs. DAF comprises two processes, namely the *inter-graph process* and the

 ISBN: 978-1-61208-203-5

intra-graph process. The inter-graph process selects, in each iteration step, a maximal common slot set and its associated MCSS-FCG to execute the intra-graph process. Its selection of the maximal common slot set is critical to meet the traffic demand. The intra-graph process assigns time slots in the selected MCSS-FCG with an objective to balance between QoS and fairness.

We note that, for a real implementation in a distributed manner, the computational complexity of obtaining a complete MCSS-FCG can be large. To reduce the control overhead and computational complexity, a MCSS-FCG used for computation may cover only the maximal cliques where the underlying flow resides. Once the computation is finished, the source/destination of the underlying flow can advertise its assignment results to the source/destination of a contending flow. To ensure a feasible time slot assignment, the final time slot assignment of a flow can be set to the minimum among all assignments of this flow suggested by the sources/destinations of this flow and contending flows. Such an assignment computed in a distributed manner could be sub-optimal. Nevertheless, we proceed in the following to describe the algorithm in its centralized form.

A. Inter-Graph Process

The following notations are used:

- Denote by $\{G_n\}$, $n = 1, 2, ..., N$, the set of MCSS-FCGs. Denote by $\{cl_n^k\}$, $k = 1, 2, ..., K$, the set of maximal cliques in G_n. Denote by d_n^k the degree of cl_n^k.
- Denote by $\{s_l\}$, $l = 1, 2, ..., L$, the set of flows in $\{G_n\}$. Denote by q_l the outstanding traffic demand of s_l.

Algorithm 1: Inter-Graph Process

```
For each G_n, set the maximal degree d̂_n of G_n to be d̂_n = max_{k=1}^{K} d_n^k;
Sort {G_n} in an ascending order of {d̂_n};
Denote by n_i the original index of the i^th MCSS-FCG in the sorted set;
for i = 1, 2, ..., N do
    Execute the intra-graph process based on G_{n_i}. Denote by q_l^{n_i} the traffic demand served in G_{n_i} for s_l after the intra-graph process;
    for l = 1, 2, ..., L do
        q_l := q_l − q_l^{n_i};
        Remove s_l from {G_n} if q_l == 0;
    end
    Update {d̂_n}. Sort {G_n} in an ascending order of {d̂_n};
end
```

The inter-graph process is specified by Algorithm 1. The process utilizes the time slots of a MCSS-FCG ahead of other MCSS-FCGs, if its flow contention level is the lowest (i.e., the highest degree of all of its maximal cliques is the lowest compared to other MCSS-FCGs). By doing this, the traffic demand of a flow can be served as much as possible before this flow enters the competition for the precious time slot resources with other contending flows in those high contention level MCSS-FCGs. Once time slots are assigned to a flow in an intra-graph process, the demand of this flow is reduced accordingly. If the demand of a flow is fully met, the flow is removed from the set of MCSS-FCGs. Hence, the inter-graph process aggressively reduces the flow contention levels of MCSS-FCGs in each iteration.

The computational complexity of the inter-graph process is identified to be $O\left(NW + N^2L + N^3\right)$, where W denotes the complexity of intra-graph process, N denotes the number of MCSS-FCGs, and L denotes the number of flows in the network. The result comes with the worst case assumption for the sorting complexity known as $O\left(k^2\right)$, where k is the number of elements for sorting. The derivation of the complexity is straightforward and hence omitted in this paper due to the page limit.

B. Intra-Graph Process

The intra-graph process iterates over maximal cliques in a MCSS-FCG. For each maximal clique, an intra-graph resource allocation algorithm calculates the number of slots to be assigned to each flow within the maximal clique. If a flow resides in multiple maximal cliques, the number of time slots assigned to the flow in this MCSS-FCG is set to the minimum of all values assigned to the flow.

The trade-off between meeting the traffic demand and preserving fairness takes the center stage of the resource allocation algorithm. The intra-graph resource allocation algorithm executed over a maximal clique strives to achieve all objectives listed at the beginning of Section III. The algorithm minimizes the total cost incurred from inadequate serving of traffic demand and allocating time slots beyond a fair share.

The cost functions associated with inadequate serving of traffic demand and allocating time slots beyond a fair share can be quite general as long as they have the following properties.

Denote by $u(z)$ the cost function induced by allocating z time slots above the fair share of a flow. With respect to a maximal clique, denote by x_i and f_i the actual number of time slots allocated to flow i and the fair share of flow i, respectively. We need

- $u(z_i) = u(x_i - f_i)$ and $u(x_i - f_i) = 0, \forall x_i \leq f_i$.
- $u(z)$ is a strictly convex and strictly increasing function w.r.t. z.

Hence, the total cost induced by allocating time slots above fair shares is $\sum_i u(x_i - f_i)I_{x_i \geq f_i}$, where I is an indicator function. $I_{x_i \geq f_i} = 1$ if $x_i \geq f_i$, otherwise $I_{x_i \geq f_i} = 0$.

Denote by $v(z)$ the cost function induced by inadequate serving of traffic demand of a flow, where z denotes the traffic demand that is not served after the allocation. Denote by R_i the data rate (in bits per slot) that can be achieved for the transmission of flow i. Denote by q_i the traffic demand (in bits per frame) from flow i. We need

- $v(z_i) = v(q_i - R_i x_i)$ and $v(q_i - R_i x_i) = 0, \forall \frac{q_i}{R_i} \leq x_i$.
- $v(z)$ is a strictly convex and strictly increasing function w.r.t. z.

Hence, the total cost induced by inadequate serving of traffic demand is $\sum_i v(q_i - R_i x_i) I_{\frac{q_i}{R_i} \geq x_i}$.

 ISBN: 978-1-61208-203-5

The optimization problem for achieving all objectives specified in Section III is described below:

$$\min_{x_1,x_2,\ldots,x_L} w_u \sum_{i=1}^{L} u(x_i - f_i) I_{x_i \geq f_i} + w_v \sum_{i=1}^{L} v(q_i - R_i x_i) I_{\frac{q_i}{R_i} \geq x_i}. \quad (1)$$

The problem is subject to the following constraints:

- The portion of traffic demand within the fair share of a flow should be fully met.

$$R_i x_i = q_i, \forall \frac{q_i}{R_i} \leq f_i. \quad (2)$$

- The actual assigned resource is no greater than q_i.

$$R_i f_i \leq R_i x_i \leq q_i, \forall \frac{q_i}{R_i} \geq f_i. \quad (3)$$

- The total number of slots assigned in a maximal clique is no larger than s_n, where s_n is the total number time slots in MCSS-FCG G_n.

$$\sum_{i}^{L} x_i \leq s_n. \quad (4)$$

The cost functions are weighted by w_u and w_v, which can be used as preferences given to QoS and fairness, respectively.

To help solving the optimization problem 1, the following derived functions are introduced.

- Denote by $U_i(x)$ the rate of cost increase induced by allocating time slots beyond fair share for flow i. Precisely, we define

$$U_i(x) = w_u \frac{\partial u(x - f_i)}{\partial x}. \quad (5)$$

  Note that, U_i increases with respect to x, since $\frac{\partial U_i(x)}{\partial x} = w_u \frac{\partial^2 u(x-f_i)}{\partial x^2} > 0$.
- Denote by $V_i(x)$ the rate of cost decrease induced by inadequately serving traffic demand for flow i. Precisely, we define

$$V_i(x) = -w_v \frac{\partial v(q_i - R_i x)}{\partial x}. \quad (6)$$

  Note that, V_i decreases with respect to x, since $\frac{\partial V_i(x)}{\partial x} = -w_v \frac{\partial^2 v(q_i - R_i x)}{\partial x^2} < 0$.

Hence, $V_i(x) - U_i(x)$ is a non-increasing function with respect to x. We call $V_i(x) - U_i(x)$ the characteristic function of flow i.

The intra-graph resource allocation process is specified by Algorithm 2. The algorithm essentially does the following: 1) when the rate of cost decrease in inadequately serving traffic demand is less than the rate of cost increase in allocating extra time slots beyond fair share, the assignment moves towards the fair share; 2) when the rate of cost decrease in inadequately serving traffic demand is more than the rate of cost increase in allocating extra time slots beyond fair share, the assignment moves towards meeting the traffic demand of the flow; 3) the number of time slots assigned to a flow is set to a value between the fair share and the traffic demand such that the assignment balances these two conflicting costs.

Algorithm 2: Intra-Graph Algorithm

for *any flow i that has $q_i \leq R_i f_i$* **do**
  Set $\hat{x_i} = \frac{q_i}{R_i}$ as the number of time slots to be assigned;
end
for *any flow i that has $q_i > R_i f_i$* **do**
  Calculate the max value and the min value of $V_i(x) - U_i(x)$. In fact, we have $V_i(f_i) - U_i(f_i) \geq V_i(x) - U_i(x) \geq V_i(\frac{q_i}{R_i}) - U_i(\frac{q_i}{R_i})$;
end
Put all max and min values of $V_i(x) - U_i(x)$ into one set Φ. Sort the elements of Φ in an increasing order. Denote the sequence by ϕ_j, $j = 1, 2, ..., J$;
for $j = 1, 2, ..., J$ **do**
  for *all flow $i = 1, 2, ...\widetilde{L}$ that do not have final slot assignments* **do**
    Calculate x_i as follows:

$$x_i = \begin{cases} f_i & \phi_j \geq V_i(f_i) - U_i(f_i) \\ \frac{q_i}{R_i} & \phi_j \leq V_i(\frac{q_i}{R_i}) - U_i(\frac{q_i}{R_i}) \\ x | V_i(x) - U_i(x) = \phi_j & \text{o.w.} \end{cases}$$

  end
  if $\sum_{i}^{L} x_i \leq s_n$ **then**
    Set $\phi_H = \phi_j$ and $\phi_L = \phi_{j-1}$;
    Break;
  end
end
for *all flow $\Omega = \{i\}$ that have $\phi_L \leq V_i(x_i) - U_i(x_i) \leq \phi_H$* **do**
  Solve the equation array $V_i(x_i) - U_i(x_i) = V_j(x_j) - U_j(x_j)$, $i \in \Omega$ and $\sum_{l=1}^{L} x_l = s_n$, L is the total number of flows;
end
Set $\hat{x_i} = x_i$;

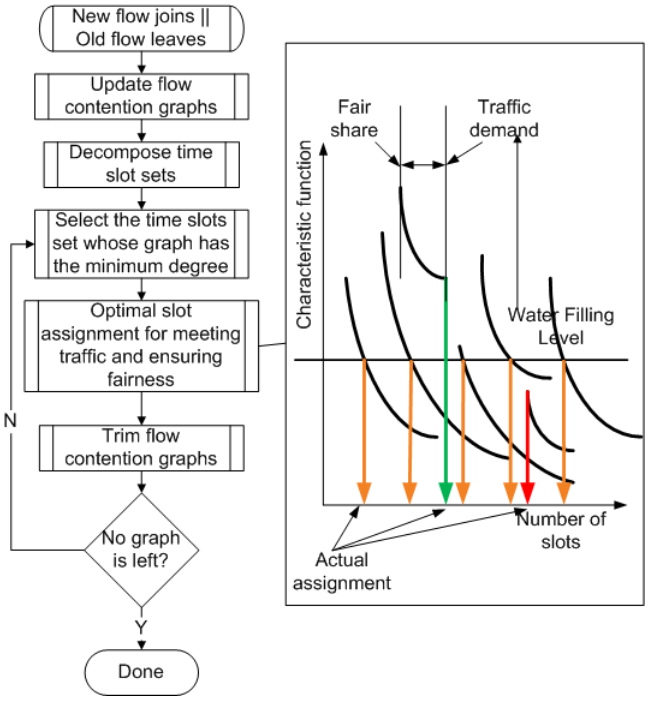


Fig. 6. DAF algorithm

Solving the equation array at Step 15 in Algorithm 2 gives the precise optimal solution, but can be computationally expensive, as $V(.)$ and $U(.)$ can be nonlinear polynomials as defined earlier. Hence, the bisection method can be used to find a solution arbitrarily close to the optimal solution with much less complexity. At Step 15 in Algorithm 2, the following bisectional iteration should be executed. In each iteration step, set $\hat{\phi} = \frac{\phi_L+\phi_H}{2}$ and set $x_i = \{x|V_i(x) - U_i(x) = \hat{\phi}\}$. If $\sum_{l=1}^{L} x_l < s_n$, set $\phi_H = \hat{\phi}$. If $\sum_{l=1}^{L} x_l > s_n$, set $\phi_L = \hat{\phi}$. And then go to the next iteration. The iteration ends when the difference between $\sum_{l=1}^{L} x_l$ and s_n is less than a negligible error margin.

A graphic representation of the resource allocation results of the intra-graph algorithm is shown in Fig. 6. The intra-graph algorithm is shown as one step in the inter-graph process. The characteristic function of each flow is evaluated against a common water filling level, which is fundamentally determined by the network scenario. A flow should be served exactly an amount so that its associated characteristic function level is closet to the common water filling level.

In the following, we prove that Algorithm 2 solves the optimization problem in Eq.1.

Theorem 1: For any functions $u(z)$ and $v(z)$ that possess the general properties described above, the optimization problem 1 has the following optimal solution:

$$\hat{x_i} = \begin{cases} \frac{q_i}{R_i} & \text{if } q_i \le R_i f_i \\ f_i & \text{if } q_i > R_i f_i \text{ and } \gamma > V_i(f_i) - U_i(f_i) \\ \in \left(f_i, \frac{q_i}{R_i}\right) & \text{if } q_i > R_i f_i \text{ and } \gamma = V_i(\hat{x_i}) - U_i(\hat{x_i}) \\ \frac{q_i}{R_i} & \text{if } q_i > R_i f_i \text{ and } \gamma < V_i(\frac{q_i}{R_i}) - U_i(\frac{q_i}{R_i}) \end{cases} \tag{7}$$

where $\gamma > 0$ is selected so that $\sum_{i}^{L} x_i \le s_n$.

Proof: From Eq. 2, we have $\hat{x_i} = \frac{q_i}{R_i}$, $\forall q_i \le R_i f_i$. This indicates that we only need to solve the problem for $q_i > R_i f_i$, so that $\frac{q_i}{R_i} > x_i$ and $x_i > f_i$. Hence, identity functions in Eqn. 1 can be subsequently removed and the problem is reduced to the following:

$$\min_{x_1,x_2,\ldots,x_{\widetilde{L}}} w_u \sum_{i=1}^{\widetilde{L}} u(x_i - f_i) + w_v \sum_{i=1}^{\widetilde{L}} v(q_i - R_i x_i) \tag{8}$$

$$\text{s.t.} \begin{cases} f_i \le x_i \le \frac{q_i}{R_i} \\ \sum_{i=1}^{\widetilde{L}} x_i \le s_n - \sum_j \frac{q_j}{R_j}, q_j \le R_j f_j \end{cases} \tag{9}$$

Let $Q_t = \sum_j \frac{q_j}{R_j}, q_j \le R_j f_j$. We have Q_t to represent the total number of time slots assigned to those flows whose traffic demands are less than their fair shares.

Based on the Karush-Kuhn-Tucker conditions, we have

$$\begin{cases} \frac{\partial\left(w_u \sum_{i=1}^{\widetilde{L}} u(\hat{x_i}-f_i) + w_v \sum_{i=1}^{\widetilde{L}} v(q_i - R_i\hat{x_i})\right)}{\partial \hat{x_i}} \\ + \frac{\partial \beta_i(R_i\hat{x_i}-q_i)}{\partial \hat{x_i}} + \frac{\partial \sigma_i(f_i-\hat{x_i})}{\partial \hat{x_i}} + \frac{\partial \gamma\left(\sum_{i=1}^{\widetilde{L}} \hat{x_i} - (s_n - Q_t)\right)}{\partial \hat{x_i}} = 0 \\ \beta_i(R_i\hat{x_i} - q_i) = 0 \\ \sigma_i(f_i - \hat{x_i}) = 0 \\ \gamma\left(\sum_{i=1}^{\widetilde{L}} \hat{x_i} - (s_n - Q_t)\right) = 0 \\ \beta_i \ge 0, \sigma_i \ge 0, \gamma \ge 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} U_i(\hat{x_i}) - V_i(\hat{x_i}) + R_i\beta_i - \sigma_i + \gamma = 0 \\ \beta_i(R_i\hat{x_i} - q_i) = 0 \\ \sigma_i(f_i - \hat{x_i}) = 0 \\ \gamma\left(\sum_{i=1}^{\widetilde{L}} \hat{x_i} - (s_n - Q_t)\right) = 0 \\ \beta_i \ge 0, \sigma_i \ge 0, \gamma \ge 0 \end{cases}$$

For the case $\sigma_i > 0$, we have $\hat{x_i} = f_i$. Consider the following sub-cases.

1) For $\beta_i = 0$, $\hat{x_i} < \frac{q_i}{R_i}$, we have $U_i(f_i) - V_i(f_i) + \gamma = \sigma_i > 0$. Hence, $\gamma > V_i(f_i) - U_i(f_i)$.
2) For $\beta_i > 0$, we have $\hat{x_i} = \frac{q_i}{R_i}$. Since we have $q_i > R_i f_i$, and $\hat{x_i} = f_i$ is contradictory to $\hat{x_i} = \frac{q_i}{R_i}$, this case is abandoned.

For the case $\sigma_i = 0$, $\hat{x_i} > f_i$, consider the following sub-cases.

1) For $\beta_i > 0$, we have $\hat{x_i} = \frac{q_i}{R_i}$ and $U_i(\frac{q_i}{R_i}) - V_i(\frac{q_i}{R_i}) + \gamma = -R_i\beta_i < 0$. Hence, we have $\gamma < V_i(\frac{q_i}{R_i}) - U_i(\frac{q_i}{R_i})$.
2) For $\beta_i = 0$, $\hat{x_i} < \frac{q_i}{R_i}$, we have $U_i(\hat{x_i}) - V_i(\hat{x_i}) + \gamma = 0$. Hence, $\gamma = V_i(\hat{x_i}) - U_i(\hat{x_i})$ and $f_i \le \hat{x_i} \le \frac{q_i}{R_i}$.

■

Theorem 2: Algorithm 2 achieves the optimal solution for the optimization problem in Eq.1.

Proof: Assume $\sum_{l=1}^{L} \frac{q_l}{R_l} > s_n$. Otherwise, the optimal solution is $\hat{x_l} = \frac{q_l}{R_l}$.

At the end of step 13 in Algorithm 2, we have identified two bounds ϕ_L and ϕ_H. Due to the execution of Step 10 and 11, setting $\gamma = \phi_L$ and setting $\hat{x_l}$ according to Eq. 7, we will have $\sum_{l=1}^{L} x_l > s_n$. Setting $\gamma = \phi_H$ and setting $\hat{x_l}$ according to Eq. 7, we will have $\sum_{l=1}^{L} x_l \le s_n$. Hence, the optimal solution exists for $\gamma \in [\phi_L, \phi_H]$.

Consider those flows $\{i\}$ where $V_i(f_i) - U_i(f_i) < \phi_L$. Since $\phi_L \le \gamma$, we have $V_i(f_i) - U_i(f_i) < \gamma$. At the end of Step 13, we already set $\hat{x_i} = f_i$, which is the optimal solution of these flows according to Theorem 1. Consider those flows where $V_j(\frac{q_j}{R_j}) - U_j(\frac{q_j}{R_j}) > \phi_H$. Since $\phi_H \ge \gamma$, we have $V_j(\frac{q_j}{R_j}) - U_j(\frac{q_j}{R_j}) > \gamma$. At the end of Step 13, we already set $\hat{x_j} = \frac{q_j}{R_j}$, which is the optimal solution of these flows according to Theorem 1. The remaining flows have $V_k(x_k) -$

$U_k(x_k) = \gamma$, which is the optimal solution for them according to Theorem 1.

■

Assume the bisection method to be used at Step 15. The computational complexity of the intra-graph process is $O\left(L^2 + L\log_2 \frac{\phi_H - \phi_L}{\Delta}\right)$, where L denotes the number of flows in the MCSS-FCG, and Δ denotes the negligible error margin. The derivation of the complexity is straightforward and hence omitted in this paper due to the page limit.

IV. Numerical Results

In this section, the performance of DAF is demonstrated through numerical experiments. The benefit of the inter-graph process is demonstrated by Scenario 1, where MCSS-FCGs are simple so that the algorithm execution order over the list of MCSS-FCGs has a more significant impact on the time slot assignment than resolving the flow contention in each MCSS-FCG. The benefit of the intra-graph process is demonstrated by Scenario 2, where MCSS-FCGs are complex so that it is critical to address the trade-off between meeting traffic demands and preserving fairness within each MCSS-FCG.

We compare the time slot allocation results obtained by DAF to those obtained by DRAND [4]. Furthermore, we extend DRAND to an enhanced version that employs our inter-graph process. We refer to the enhanced version of DRAND as E-DRAND. E-DRAND selects, in each iteration step, a maximal common slot set and its associated MCSS-FCG to execute the generic DRAND algorithm. E-DRAND helps identify the benefit of the inter-graph process. Numerical results show that DAF performs significantly better than DRAND and E-DRAND.

A. Simulation Model

The inputs of our program include flows, their traffic demands, and MCSS-FCGs, so that the program does not implement control protocols or graph computation algorithms to prepare these input parameters. We did not conduct our simulation using a full fledged network simulator, such as NS-2, since a generic program can provide us rich scenarios with flexible input parameters, without the involvement of inessential network events in a full fledged network simulator.

The example function $u(z)$ that represents the cost induced by allocating extra time slots above the fair share is set to $u(z) = z$, $z \geq 0$, equivalent to $u(x - f) = x - f, x \geq f$. Hence, we have $U_i(x) = w_u$. The example function $v(z)$ that represents the cost induced by inadequately serving traffic demand is set to $v(z) = \frac{1}{q-z} - \frac{1}{q}$, $0 \leq z \leq q$, equivalent to $v(q - Rx) = \frac{1}{Rx} - \frac{1}{q}$, $0 \leq x \leq \frac{q}{R}$. Hence, we have $V(x) = \frac{w_v}{Rx^2}$. Under this setting, the cost induced by inadequately serving traffic demand increases drastically when more traffic demand is not served.

We set the weights for these cost functions so that their rate of changes are about the same. Assume a typical flow requires 50 time slots among 256 available time slots. We set $\frac{w_u}{w_v} = \frac{1}{2500R}$. For a concise presentation, we assume a common data rate $R = 1$ bit per slot for all flows.

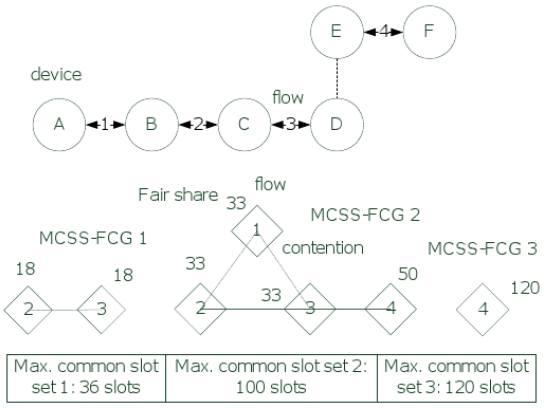


Fig. 7. Network Scenario 1

The criterion for fairness is defined with low computational complexity. The fair share for a flow i, in a MCSS-FCG, G_n, is set to be equal to $\frac{s_n}{d_i^n}$, where d_i^n is the maximum degree of all maximal cliques flow i resides within G_n, s_n is the number of time slots available in G_n. An assignment is said to be fair for flows in the MCSS-FCG, if every flow in the MCSS-FCG is assigned an amount of slots that is larger than or equal to its fair share.

To quantitatively evaluate the performance of the algorithms, we compare the values of the objective function specified in Eq. 1 calculated on the overall contention graph. The objective function specified in Eq. 1 represents the total cost incurred from inadequate serving of traffic demand and allocating time slots beyond a fair share.

B. Network Scenario 1

A network scenario with a relatively low flow contention level is studied to highlight the benefit of the inter-graph process. In this setup, fairness is a lesser issue, since traffic demand can be well served by using a particular order in which the inter-graph process selects MCSS-FCGs. The scenario is illustrated in Fig. 7.

There are 4 new flow arrivals. Each flow has an opportunity to utilize some time slots. Three maximal common slot sets and their corresponding MCSS-FCGs are identified. Flow 3 can utilize 136 time slots in total (36 from MCSS 1 and 100 from MCSS 2), however it has to contend with many other flows in two maximal common slot sets. Flow 4 can utilize 120 time slots without facing any contention from other flows. Two traffic demand patterns are simulated. They are specified by their corresponding traffic demand vectors in Table I and Table II. Traffic pattern 2 has slightly higher traffic demand than traffic pattern 1. The results under traffic pattern 1 and traffic pattern 2 are shown in Table I and Table II, respectively.

DAF is shown to meet the joint QoS and fairness requirement much better than DRAND and perform about the same as E-DRAND does, as it achieves the lowest objective function value. This result highlights the significant benefit of using our inter-graph process.

 ISBN: 978-1-61208-203-5

TABLE I
SLOT ASSIGNMENT COMPARISON UNDER TRAFFIC PATTERN 1

Flow ID	Demand (slots)	DAF	DRAND	E-DRAND
#1	35	35	28	35
#2	70	52	46	65
#3	35	35	35	35
#4	100	100	100	100
	Objective Value	-70.6374	-59.5093	-67.2527

TABLE II
SLOT ASSIGNMENT COMPARISON UNDER TRAFFIC PATTERN 2

Flow ID	Demand (slots)	DAF	DRAND	E-DRAND
#1	50	36	24	36
#2	50	50	42	50
#3	50	50	46	50
#4	100	100	100	100
	Objective Value	-49.5556	-24.9617	-49.5556

TABLE III
SLOT ASSIGNMENT DETAILS UNDER TRAFFIC PATTERN 1

Flow ID	FCG #1	FCG #2	FCG #3
#1		35	
#2	18	35	
#3	18	17	
#4			100

The time slot assignment in each MCSS-FCG under traffic pattern 1 is shown in Table III. The following observations can be drawn from traffic pattern 1's results:

- The advantage of the inter-graph process can be clearly seen. Flow 4 is assigned as many slots as possible in G_3 due to the effect of the inter-graph process, so that its high demand does not interfere with other flows.
- Traffic demands of flow 1 and 3 are completely served due to their relatively low traffic demand levels. Flow 1's assignment in G_2 is slightly above its fair share (33 slots) since assignment for flow 2 and 3 in G_1 has relieved the contention in G_2. Again, the advantage of the inter-graph process is evident.
- Time slots of G_2 are not all utilized, since a flow only accepts the minimum of all assignments it receives from all maximal cliques within a graph.

The time slot assignment in each MCSS-FCG under traffic pattern 2 is shown in Table IV. The following observations can be drawn from traffic pattern 2's results:

- Traffic demand of flow 1 cannot be fully met due to its high level of traffic demand in the only MCSS-FCG, G_2, it resides in. Note that its assignment is above its fair share.
- Flow 2 and 3 are in a topologically similar position, as the degrees of their maximal cliques are equal in G_1 and G_2 and they contend with similar set of flows. When traffic demand levels of flow 2 and 3 become more even compared to theirs in traffic pattern 1, their traffic demand can be fully met.

TABLE IV
SLOT ASSIGNMENT DETAILS UNDER TRAFFIC PATTERN 2

Flow ID	FCG #1	FCG #2	FCG #3
#1		36	
#2	18	32	
#3	18	32	
#4			100

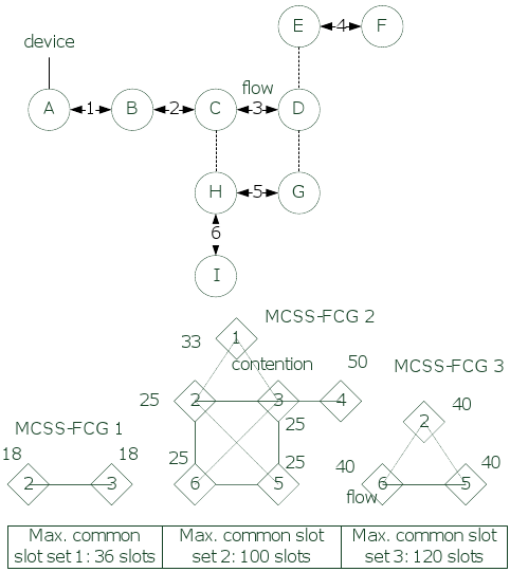


Fig. 8. Network Scenario 2

C. Network Scenario 2

A network scenario with a higher flow contention level is studied to highlight the benefit of the intra-graph process. In this setup, more fairness problems would need to be addressed by the intra-graph process and algorithm. The scenario is illustrated in Fig. 8.

There are 6 new flow arrivals. Three maximal common slot sets and their MCSS-FCGs are identified. Flow 2, 3, 5 and 6 are in a high contention level among each other in various MCSS-FCGS $\{G_1, G_2, G3\}$. Two traffic demand patterns are simulated. They are specified by their corresponding traffic demand vectors in Table V and Table VI. Traffic pattern 4 has slightly higher traffic demand than traffic pattern 3. The results under traffic pattern 3 and traffic pattern 4 are shown in Table V and Table VI, respectively.

DAF is shown to meet the joint QoS and fairness requirement much better compared to DRAND and E-RAND, as it achieves the lowest objective function value. This result highlights the significant benefit of using our intra-graph process.

The time slot assignment in each MCSS-FCG under traffic pattern 3 is shown in Table VII. The following observations can be drawn from traffic pattern 3's results:

- All traffic demand requirements are satisfied. This is due

TABLE V
SLOT ASSIGNMENT COMPARISON UNDER TRAFFIC PATTERN 3

Flow ID	Demand (slots)	DAF	DRAND	E-DRAND
#1	50	50	17	22
#2	50	50	50	50
#3	50	50	35	39
#4	50	50	16	21
#5	50	50	50	50
#6	50	50	50	50
	Objective Value	-39.0000	103.7374	39.7865

TABLE VI
SLOT ASSIGNMENT COMPARISON UNDER TRAFFIC PATTERN 4

Flow ID	Demand (slots)	DAF	DRAND	E-DRAND
#1	60	42	17	19
#2	60	60	60	60
#3	60	60	35	37
#4	60	50	16	20
#5	60	60	60	60
#6	60	60	60	60
	Objective Value	19.1905	158.7374	116.146

TABLE VII
SLOT ASSIGNMENT DETAILS UNDER TRAFFIC PATTERN 3

Flow ID	FCG #1	FCG #2	FCG #3
#1		50	
#2	18		32
#3	18	32	
#4		50	
#5		10	40
#6		10	40

to the fact that G_1 and then G_3 are able to sequentially provide large amount of time slots for flow 2, 5, and 6, which leads to a significantly lowered traffic demand in a highly contended MCSS-FCG G_2. As an example, flow 2 does not need to be assigned any time slots in G_2.

The time slot assignment in each MCSS-FCG under traffic pattern 4 is shown in Table VIII. The following observations can be drawn from traffic pattern 4's results:

- Flow 1's traffic demand cannot be fully accommodated in G_2, since its contending flow, flow 3 has an increased its demand. Note that flow 1's assignment in G_2 is surely still above its fair share.
- The tradeoff between meeting traffic demand and maintaining fairness thereby reducing potential congestion is addressed by DAF. The time slots of G_2 are not all utilized, although flow 1 still has traffic demand to meet. This is because many flows have been allocated time slots beyond their fair shares.

TABLE VIII
SLOT ASSIGNMENT DETAILS UNDER TRAFFIC PATTERN 4

Flow ID	FCG #1	FCG #2	FCG #3
#1		42	
#2	18	2	40
#3	18	42	
#4		50	
#5		20	40
#6		20	40

V. CONCLUSIONS

In this paper, a demand-aware fair resource allocation algorithm is proposed to allocate time slots in TDMA-based multi-hop wireless networks with the objective of meeting both traffic demand as much as possible while enforcing a predefined fairness level. The algorithm projects the new network flow arrivals onto multiple Maximal Common Slot Set based Flow Contention Graphs and then executes an intra-graph resource allocation algorithm over contention graphs in a carefully selected order. The execution order strives to reduce the flow contention as much as possible before the intra-graph algorithm starts allocating time slots over each particular graph. The proposed intra-graph algorithm is proven to minimize, over a maximal clique, a generic cost function value incurred when inadequately serving traffic demand and serving beyond fair shares. Numerical experiments are conducted to demonstrate the effectiveness of the proposed algorithm. The proposed algorithm is shown to well meet the traffic demand and achieve the predefined fairness.

REFERENCES

[1] L. B. Jiang, and S. C. Liew, *Proportional fairness in wireless LANs and ad hoc networks*, Proceedings of IEEE WCNC 3, volume 3, pp. 1551–1556, March 2005.

[2] X. L. H., and B. Bensaou, *On Max-min Fairness Bandwidth Allocation and Scheduling in Wireless Ad Hoc Networks: Analytical Framework and Implementation*, Proceedings of ACM MobiHoc, pp. 221–231, 2001.

[3] L. Chen, S. H. Low, and J. C. Doyle, *Joint congestion control and media access control design for ad hoc wireless networks*, Proceedings of IEEE INFOCOM 3, volume 3, pp. 2212–2222, March 2005.

[4] I. Rhee, A. Warrier, and J. Min: DRAND, *Distributed Randomized TDMA Scheduling for Wireless Ad-hoc Networks*, Proceedings of ACM MobiHoc, pp. 190–201, 2006.

[5] H. Zhai, *QoS Support over UWB Mesh Networks*, Proceedings of IEEE WCNC, pp. 2283–2288, March 2008.

[6] *Standard ECMA-368 High Rate Ultra Wideband PHY and MAC Standard*, http://www.ecma-international.org/publications/standards/Ecma-368.htm, December 2008.

[7] *Standard ECMA-387 High Rate 60 GHz PHY, MAC and HDMI PAL*, http://www.ecma-international.org/publications/standards/Ecma-387.htm, December 2008.

[8] A. M. Chou, and V. O. K. Li, *Slot Allocation Strategies for TDMA Protocols in Multihop Packet Radio Networks*, Proceedings of IEEE INFOCOM 2, volume 2, pp. 710–716, May 1992.

[9] J. Grnkvist, *Traffic Controlled Spatial Reuse TDMA in Multi-hop Radio Networks*, Proceedings of IEEE PIMRC 3, volume 3, pp. 1203–1207, September 1998.

[10] P. Bjrklund, Vrbrand, P., and Yuan, D., *Resource Optimization of Spatial TDMA in Ad Hoc Radio Networks: A Column Generation Approach*, Proceedings of IEEE INFOCOM 2, volume 2, pp. 818–824, April 2003.

[11] H. L. Chao, and W. Liao, *Credit-Based Slot Allocation for Multimedia Mobile Ad Hoc Networks*, IEEE Journal on Selected Areas in Communications 21(10), volume 21, pp. 1642–1651, December 2003.

[12] T. Nandagopal, T. E. Kim, X. Gao, and V. Bharghavan, *Achieving MAC layer fairness in wireless packet networks*, Proc. ACM MobiCom, pp. 87–98, September 2000.

 ISBN: 978-1-61208-203-5

An Analysis of the Interference Problem in Wireless TDMA Networks

Anuschka Igel and Reinhard Gotzhein
Networked Systems Group
University of Kaiserslautern, Germany
{igel,gotzhein}@cs.uni-kl.de

***Abstract*—Communication in wireless networks raises the so-called *interference problem*, which means that the transfer of a message from some node *a* to a receiving node *b* can be disturbed by the overlapping transmission of another node *c* in interference range of node *b*. There are several approaches to solve this problem, such as the exclusive reservation of network-wide synchronized time slots in interference range of both sender and receiver, which we formalize and study in this paper. We first show that the interference problem can be solved if each node knows the current communication and interference topology of the network and the transmission reservations of all nodes at any time, and if reservations take place in a coordinated manner. We then analyze how far this global status information can be reduced while preserving the solvability of the interference problem. We apply our findings to evaluate some existing reservation protocols concerning their abilities to solve the interference problem, and identify possible shortcomings.**

Keywords*—*interference problem; reservation; TDMA; neighborhood; wireless network.

I. INTRODUCTION

Wireless networks are a commonly used technology these days. Basic problems of wireless networks, such as varying channel quality, interference due to concurrent transmissions, and energy shortage, have been addressed by a variety of sophisticated approaches to channel coding and medium arbitration. However, today's prevailing contention-based medium access techniques are highly prone to frame collisions when applied in multi-hop networks, due to the interference problem, which is illustrated in Figure 1. The figure shows the topology and a scenario of a wireless multi-hop network. The topology distinguishes communication links for data transfer, and interference links that may prevent successful data transfer if used concurrently. For simplicity, we assume that in this network, all nodes use the same frequency and code. In the scenario, nodes a and b want to exchange a message m. For a successful transfer, it is necessary but not sufficient that all nodes in *communication range* of node b (except a) stay silent while m is being transmitted. If, e. g., a transmission of node c, which is in *interference range* of node b, overlaps with a's transmission, the transfer would fail.

To solve the interference problem, a variety of exclusive reservation schemes using TDMA (*Time Division Multiple Access*), FDMA (*Frequency Division Multiple Access*), CDMA (*Code Division Multiple Access*), SDMA (*Space Division Multiple Access*) or combinations thereof are conceivable (see [1]).

In this paper, we study exclusive reservation schemes based on TDMA, where time slots are synchronized network-wide with an upper bound for clock offset. A synchronization protocol with this property has been published in [2]. To improve bandwidth usage, we additionally consider SDMA. We stipulate that if two nodes a and b want to communicate, they must reserve a free time slot s exclusively in interference range of a and b. More precisely, this means that s is not yet reserved for reception by any other node in interference range of a, nor for transmission by any other node in interference range of b. It is obvious that by always following these reservation rules, overlapping transmissions are safely avoided, and the interference problem is solved.

In this paper, we formalize the interference problem and define a global reservation criterion that builds on complete status information to solve the problem in a TDMA / SDMA setting. Because this criterion is too expensive to be implemented, we then examine how far the complete status information can be reduced while still solving the interference problem. This, finally, leads to a local reservation criterion based on a reduced network view and derived localized reservation status predicates, which we prove to be equivalent to the global criterion. Finally, we assess existing reservation protocols concerning their abilities to solve the interference problem, and identify possible shortcomings. In our future work, we plan to devise efficient reservation protocols based on the local reservation criterion, with nodes learning about their relevant reservation status by simply observing reservation traffic.

This paper is organized as follows: In Section II, we formally define our network model. The global reservation criterion is defined in Section III. In Section IV, we provide an equivalent local reservation criterion that is based on a reduced network view. In Section V, we analyze existing reservation protocols, discuss related work in Section VI, and draw conclusions in Section VII.

II. NETWORK MODEL

We now introduce our network model, which distinguishes between (wireless) communication and interference links. We say that a node a is in *communication range* of a node b if a transmission of a is received correctly by b. Node a is in *interference range* of b if a transmission of a can prevent the correct reception of a concurrent transmission of some other node c to b.

 ISBN: 978-1-61208-203-5

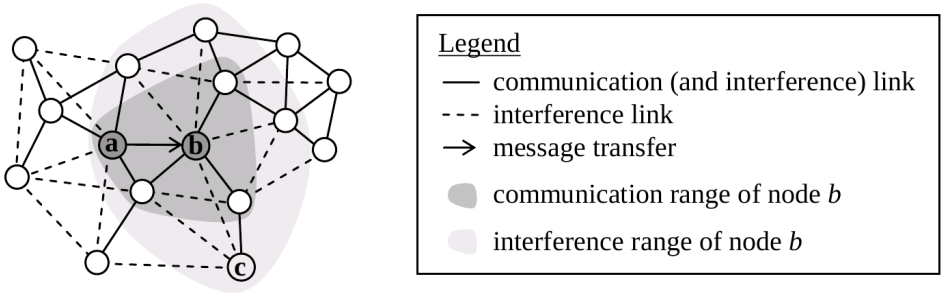


Figure 1. Illustration of the interference problem.

Definition 1. *Let V be a set of nodes. Then* (wireless) communication *and* interference links *are formally expressed by the following relations:*

- $CL =_{Df}$ $\{(a,b) \in V \times V : a$ *is in communication range of* $b\}$
- $IL =_{Df}$ $\{(a,b) \in V \times V : a$ *is in interference range of* $b\}$

We assume that there is only one antenna per node, which means that a node can neither receive nor detect any interference while transmitting. Furthermore, we assume that all links are bidirectional, which can be achieved in practice by taking suitable detection measures. These assumptions are formalized by requiring CL and IL to be irreflexive and symmetric. In addition, $CL \subseteq IL$ holds (we call this *consistency criterion*).

Based on the relations CL and IL, we define our model of a (wireless) network, which is a graph with two kinds of edges representing communication and interference links, and where all pairs of nodes are connected through a path of communication links.

Definition 2. *Let V be a set of nodes, CL and IL be relations expressing communication and interference links, respectively. A* (wireless) network *is formally modeled as a directed graph $G = (V, L, E)$, where $L = \{cl, il\}$ is a set of* labels*, and $E \subseteq V \times V \times L$ is a set of* edges*. The set $E = E_{cl} \cup E_{il}$ is composed of the following subsets:*

- $E_{cl} =_{Df} \{(a,b,l) \in V \times V \times L : l = cl \wedge CL(a,b)\}$ *(communication links)*
- $E_{il} =_{Df} \{(a,b,l) \in V \times V \times L : l = il \wedge IL(a,b)\}$ *(interference links)*

In addition, the communication subgraph $G_{cl} =_{Df} (V, \{cl\}, E_{cl})$ has to be connected, i. e., $\forall a, b \in V : \exists p =_{Df} (v_1^p, \ldots, v_{|p|+1}^p) \in V^+$ such that $\forall i \in \{1, \ldots, |p|\} : (v_i^p, v_{i+1}^p, cl) \in E_{cl}$, $v_1^p = a$ and $v_{|p|+1}^p = b$. We require that p is a cycle-free path*, i. e., no node occurs more than once in p. The length $|p|$ of p is its number of edges. The* communication distance *between two nodes $a, b \in V$ is defined as*

$$d_{G_{cl}}(a,b) =_{Df} \min_{p \in P_{G_{cl}}(a,b)} |p| \,,$$

where $P_{G_{cl}}(a,b)$ is the set of all cycle-free communication paths starting in a and ending in b. The interference distance *$d_{G_{il}}$ is defined analogously.*

Since CL and IL are irreflexive, no node has a communication or interference link to itself. Since the relations are symmetric, all links are bidirectional, i. e., $\forall (a,b,l) \in E : (b,a,l) \in E$, which is equivalent to regarding an undirected graph. In the following, we write $G = (V, E)$ to refer to a network $G = (V, L, E)$, since the labeling is fixed.

We assume that the *Single Network Property* holds, which means that all nodes are connected via some path of communication links (this is already covered by Definition 2), and no other nodes in interference range that apply a different MAC protocol are active in the same frequency band. This property can be satisfied in real environments by sufficient topology control combined with standardization measures and / or frequency and spatial division. Furthermore, slot reservation is a long-term functionality, which requires a sufficiently stable network topology to prevent frequent loss of reservations due to link breaks.

Next, we introduce several notions of neighborhood between nodes:

Definition 3. *Let $G = (V, E)$ be a (wireless) network, $a \in V$ and $i \geq 0$ an integer value.*

i) The i-hop communication *and* interference neighborhoods *of a are defined as*

$$CN_i(a) =_{Df} \{b \in V : d_{G_{cl}}(a,b) = i\} \,.$$

$$IN_i(a) =_{Df} \{b \in V : d_{G_{il}}(a,b) = i\} \,.$$

ii) The maximal i-hop communication *and* interference neighborhoods *of a are defined as*

$$CN_{\leq i}(a) =_{Df} \{b \in V : d_{G_{cl}}(a,b) \leq i\} \,.$$

$$IN_{\leq i}(a) =_{Df} \{b \in V : d_{G_{il}}(a,b) \leq i\} \,.$$

In the following, we also denote 1-hop communication neighbors simply as neighbors. From the definition, it follows that the 0-hop communication / interference neighborhood of a node is the node itself. Since neighborhood is defined w. r. t. the shortest path, i-hop neighbors are not $(i+j)$-hop neighbors for any $j > 0$. However, i-hop neighbors are also *maximal* $(i+j)$-hop neighbors for every $j \geq 0$. Because of the consistency criterion, $CN_1(a) \subseteq IN_1(a)$ holds for all $a \in V$.

 ISBN: 978-1-61208-203-5

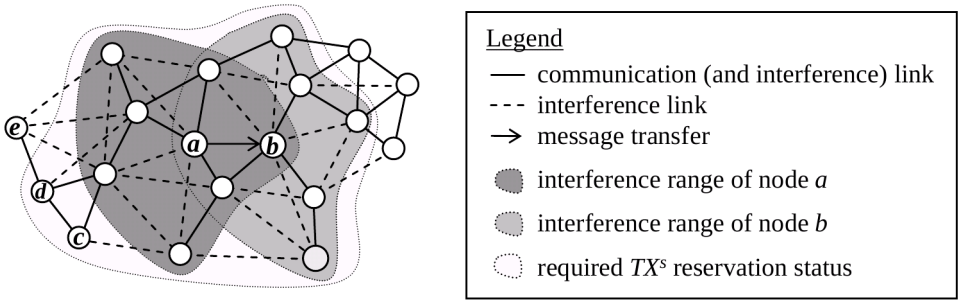


Figure 2. Illustration of the global reservation criterion.

Definition 4. *Let $G = (V, E)$ be a (wireless) network, $a \in V$, $P(a) \subseteq V$ and $R(a) \subseteq V$ be unary relations over V. Then we define*

$$P(R(a)) =_{Df} \{c \in V : \exists\, b \in R(a) : c \in P(b)\}$$

Note that for all $a \in V$ and $i \geq 0$, the sets $CN_i(a)$, $CN_{\leq i}(a)$, $IN_i(a)$ and $IN_{\leq i}(a)$ are relations according to Definition 4.

An example illustrating Definition 4 is the (1-hop) interference neighborhood of the communication neighborhood of a node a, which is denoted by $IN_1(CN_1(a))$. Note that $a \in IN_1(CN_1(a))$ holds, provided $CN_1(a) \neq \emptyset$.

III. Global Reservation Criterion

We assume that time is structured into macro slots, which are subdivided into consecutively numbered micro slots. The set of micro slots will be denoted by S in the following; the notions micro slot, time slot and slot will be used interchangeably. If a slot is reserved, then this reservation holds for all following macro slots, until it is released.

We now formally state the *global reservation criterion* $F^s_{TX}(a, b)$, defining whether a time slot $s \in S$ is free for transmissions from node a to node b, where b is in communication range of a. Informally, this means that s is currently reserved neither for reception by any node in interference range of a, nor for transmission by any node in interference range of b. The reservation criterion is called *global*, because it is based on global knowledge about network topology and reservation status.

Definition 5 (Reservation status)**.** *Let $G = (V, E)$ be a network, and $s \in S$ be a time slot. The* reservation status *$TX^s \subseteq V \times V$ of slot s defines, for all pairs of nodes $a, b \in V$, whether s is reserved for transmissions from a to b, provided $b \in CN_1(a)$. The following relations are derived from TX^s:*

- $TX^s(a) =_{Df} \exists\, b \in V : TX^s(a, b)$
  s reserved for transmission by node a
- $RX^s(a, b) =_{Df} TX^s(b, a)$
  s reserved by node a for reception from b
- $RX^s(a) =_{Df} \exists\, b \in V : RX^s(a, b)$
  s reserved for reception by node a

Please note that the derived relations do not carry any additional status information, but are introduced for better readability of the global reservation criterion:

Definition 6 (Global reservation criterion)**.** *Let $G = (V, E)$ be a network, $a, b \in V$, $b \in CN_1(a)$, $s \in S$ be a time slot, and TX^s be the reservation status of slot s. The* global reservation criterion *F^s_{TX} defines whether s is free for transmissions from a to b:*

$$F^s_{TX}(a, b) =_{Df} \forall\, c \in IN_{\leq 1}(a) :\ \neg\, RX^s(c) \wedge \forall\, d \in IN_{\leq 1}(b) :\ \neg\, TX^s(d)$$

We recall that $IN_{\leq 1}(a)$ and $IN_{\leq 1}(b)$ include nodes a and b, respectively. Therefore, the definition covers the necessary condition that both a and b have reserved s neither for transmission nor for reception. From the definition, it follows immediately that the interference problem can be solved if each node knows the current communication and interference topology of the network and the sending reservations of all nodes.

Figure 2 illustrates the global reservation criterion $F^s_{TX}(a, b)$. In the figure, all nodes whose TX^s reservation status is required to solve the interference problem are highlighted by the outer shape. This includes all nodes in interference range of a and b (inner shapes), and nodes c and d, but not node e. Nodes outside the interference range of a (e. g., c and d) have to be considered if the relation RX^s is not directly available but is derived from TX^s. If, for example, a value $RX^s(f, g)$ is needed, it is derived from $TX^s(g, f)$.

IV. Local Reservation Criteria

In this section, we transform the definition of the global reservation criterion $F^s_{TX}(a, b)$ into two *local* forms by replacing global predicates with local ones, thereby reducing the status information required to solve the interference problem. *Local predicates* are predicates that are defined from the point of view of a *single* node. We proceed in two steps: In Section IV-A, we introduce a local definition that is based on local knowledge about nodes in interference neighborhood, and show that it is equivalent to the global definition. In Section IV-B, we introduce an assumption about interference neighborhood such that only local knowledge about nodes in communication neighborhood is required to solve the interference problem, which provides a basis for feasible reservation protocols.

 ISBN: 978-1-61208-203-5

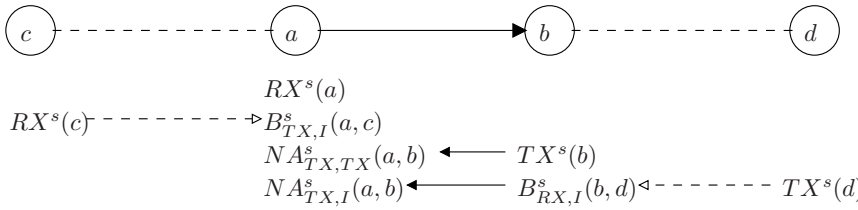


Figure 3. Topology pattern and local predicates to define the local reservation criterion with interference neighborhood.

A. Local Reservation Criterion with Interference Neighborhood

In this section, we assume that nodes have access to reservation status information of nodes in communication *and* interference neighborhood. From this status information, nodes can derive local predicates, which they can use to determine whether the reservation criterion is satisfied. Figure 3 shows a topological pattern that we use to define the local reservation criterion. Interference links are represented by dashed lines, communication links (which are also interference links) by solid lines. The arrow indicates that there is a request to reserve some slot s for transmission from a to b. To check the reservation criterion, status information expressed by local predicates that are listed beneath the nodes is required. An arrow from a predicate P to a predicate Q denotes that P is used to derive Q. A dashed arrow indicates that predicate values may not be directly available, which is the case if the corresponding node is in interference range, but not in communication range.

The global reservation criterion $F^s_{TX}(a,b)$ (see Definition 6) assumes $b \in CN_1(a)$ and is based on global predicates TX^s and RX^s:

$$\begin{aligned} F^s_{TX}(a,b) =_{Df} & \forall\, c \in IN_{\leq 1}(a) :\ \neg\, RX^s(c) \wedge \\ & \forall\, d \in IN_{\leq 1}(b) :\ \neg\, TX^s(d) \\ \equiv \neg\, RX^s(a) \ \wedge\ & \forall\, c \in IN_1(a) :\ \neg\, RX^s(c) \wedge \\ \neg\, TX^s(b) \ \wedge\ & \forall\, d \in IN_1(b) :\ \neg\, TX^s(d) \end{aligned}$$

To finally replace global predicates in this definition, we start by defining two local predicates:

Definition 7. $B^s_{TX,I}(a,b) =_{Df} b \in IN_1(a) \wedge RX^s(b)$
Slot s is blocked (B) for transmission (TX) at node a because of possible interference with a reception of b, with b in interference range (I) of a.

Definition 8. $B^s_{RX,I}(a,b) =_{Df} b \in IN_1(a) \wedge TX^s(b)$
Slot s is blocked (B) for reception (RX) at node a because of possible interference with a transmission of b, with b in interference range (I) of a.

Obviously, $TX^s(a,b)$ implies $B^s_{TX,I}(a,b)$ and $RX^s(a,b)$ implies $B^s_{RX,I}(a,b)$. However, the predicate TX^s (RX^s) carries the additional information which node is the sender (receiver) in slot s. Inserting these predicates, $F^s_{TX}(a,b)$ can be restated as:

$$\begin{aligned} F^s_{TX}(a,b) \equiv \neg\, RX^s(a) \wedge \forall c \in IN_1(a) : \neg\, B^s_{TX,I}(a,c) \wedge \\ \neg\, TX^s(b) \wedge \forall d \in IN_1(b) : \neg\, B^s_{RX,I}(b,d) \end{aligned}$$

In this definition, some predicates are local to the transmitting node a, while others are local to the receiving node b. We now define further predicates to obtain a definition of $F^s_{TX}(a,b)$ that is entirely local to a:

Definition 9. $NA^s_{TX,TX}(a,b) =_{Df} b \in CN_1(a) \wedge TX^s(b)$
Slot s is not available (NA) for transmission from a to b, because b has already reserved this slot for transmission.

Definition 10. $NA^s_{TX,I}(a,b) =_{Df} b \in CN_1(a) \wedge \exists\, c \in IN_1(b) :\ B^s_{RX,I}(b,c)$
Slot s is not available (NA) for transmission from a to b because of a possible interference with a transmission of some node c in interference range of b.

Inserting these predicates into the restated predicate $F^s_{TX}(a,b)$ above yields:

$$\begin{aligned} F^s_{TX}(a,b) \equiv \neg\, RX^s(a) \wedge \forall c \in IN_1(a) : \neg\, B^s_{TX,I}(a,c) \wedge \\ \neg\, NA^s_{TX,TX}(a,b) \wedge \neg\, NA^s_{TX,I}(a,b) \end{aligned}$$

Please note that first, this restatement of $F^s_{TX}(a,b)$ is equivalent to the definition of the global reservation criterion. Second, it is based on local knowledge about the reservation status of nodes in interference neighborhood of node a only, therefore, it is a local definition of $F^s_{TX}(a,b)$. It now remains to be shown how node a can determine its local values of these predicates.

We observe that the value of $RX^s(a)$ can directly be obtained from the list of current reservations of a. To determine the values of $B^s_{TX,I}$, the RX^s values of all *interference* neighbors of a are needed, which, however, may be out of communication range. $NA^s_{TX,TX}(a,b)$ can be derived from the TX^s values of b. Finally, to calculate $NA^s_{TX,I}(a,b)$, the $B^s_{RX,I}$ values of b are needed. For b to calculate its $B^s_{RX,I}$ values, the TX^s values of its interference neighbors are needed.

Obviously, although the above definition of $F^s_{TX}(a,b)$ is local, there still exists no reservation protocol that can solve the interference problem in the general case, as an exchange of status information with all nodes in interference neighborhood would be required.

B. Local Reservation Criterion with Communication Neighborhood

To determine the values of predicates $B^s_{TX,I}(a,c)$ and $B^s_{RX,I}(b,d)$ of nodes a and b, status information of interference neighbors c and d is needed. Since these neighbors may not be in communication range, it is not obvious how this information can

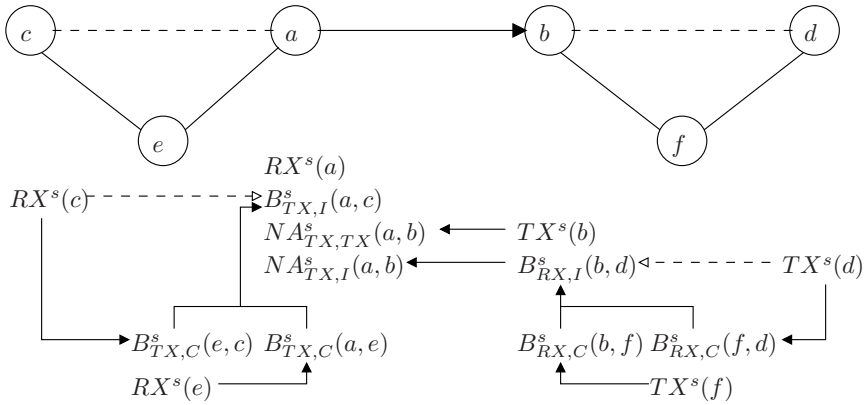


Figure 4. Topology pattern and local predicates to define the local reservation criterion with communication neighborhood, with $IN_1(a) = CN_1(a) \cup CN_2(a)$.

be acquired. Recall that in our definition of *network*, we stipulate that all nodes are connected via some path of communication links. Therefore, two nodes in interference range are always connected by a path of length $\leq n-1$, where n is the number of nodes in the network. If we additionally assume that we can control the topology to a certain extent, we may limit the communication distance of nodes in interference range to a value d, with $2 \leq d$. This means that $\forall\, a \in V : IN_{\leq 1}(a) \subseteq CN_{\leq d}(a)$ holds. For conservative decisions, we further assume that all nodes with a communication distance of at most d are in interference range, i. e., $\forall\, a \in V : IN_{\leq 1}(a) \supseteq CN_{\leq d}(a)$. On the whole, this means $\forall\, a \in V : IN_{\leq 1}(a) = CN_{\leq d}(a)$ or $\forall\, a \in V : IN_1(a) = CN_{\leq d}(a) \setminus a$, respectively.

In the following, we assume $d = 2$, which means that all nodes in interference range, but not in communication range have a communication distance of 2. This means that in the following, $\forall\, a \in V : IN_1(a) = CN_1(a) \cup CN_2(a)$ holds. Figure 4 extends the topological pattern of Figure 3, capturing this assumption and adding auxiliary predicates of nodes in communication range that are used to replace predicates of nodes in interference neighborhood. The meaning of arrows is as in Figure 3.

Definition 11. $B^s_{TX,C}(a,b) =_{Df} b \in CN_1(a) \wedge RX^s(b)$
Slot s is blocked for transmission at node a because of a possible interference with a reception of b, with b in communication range (C) of a.

Definition 12. $B^s_{RX,C}(a,b) =_{Df} b \in CN_1(a) \wedge TX^s(b)$
Slot s is blocked for reception at node a because of a possible interference with a transmission of b, with b in communication range (C) of a.

Note that the definition of $B^s_{TX,C}(a,b)$ ($B^s_{RX,C}(a,b)$) slightly differs from the definition of $B^s_{TX,I}(a,b)$ (see Definition 7) ($B^s_{RX,I}(a,b)$ (see Definition 8)), as nodes in communication range instead of interference neighborhood are considered. We remark that the formal definitions of $NA^s_{TX,TX}(a,b)$ and $B^s_{RX,C}(a,b)$ are the same. However, for conceptual clarity, we prefer to use two predicates.

Based on the assumption $IN_1(a) = CN_1(a) \cup CN_2(a)$, the predicates $B^s_{TX,I}$ and $B^s_{RX,I}$ can be derived from the reservation status of nodes in communication range as follows:

$$\begin{aligned} B^s_{TX,I}(a,b) =_{Df}\ & b \in IN_1(a) \wedge RX^s(b) \\ \equiv\ & b \in (CN_1(a) \cup CN_2(a)) \wedge RX^s(b) \\ \equiv\ & (b \in CN_1(a) \wedge RX^s(b)) \vee \\ & (b \in CN_2(a) \wedge RX^s(b)) \\ \equiv\ & B^s_{TX,C}(a,b) \vee \exists\, c \in CN_1(a) : \\ & (b \in CN_1(c) \wedge RX^s(b)) \\ \equiv\ & B^s_{TX,C}(a,b) \vee \exists\, c \in CN_1(a) : \\ & B^s_{TX,C}(c,b) \end{aligned}$$

$$\begin{aligned} B^s_{RX,I}(a,b) =_{Df}\ & b \in IN_1(a) \wedge TX^s(b) \\ \equiv\ & (b \in CN_1(a) \wedge TX^s(b)) \vee \\ & (b \in CN_2(a) \wedge TX^s(b)) \\ \equiv\ & B^s_{RX,C}(a,b) \vee \exists\, c \in CN_1(a) : \\ & B^s_{RX,C}(c,b) \end{aligned}$$

This way, a can derive the $B^s_{TX,I}$ values from its own $B^s_{TX,C}$ values and those of its neighbors (in general, the $B^s_{TX,C}$ values of all nodes in $CN_{\leq d-1}(a)$ would be needed). The $B^s_{TX,C}$ values can in turn be determined by each node from the RX^s values of its neighbor nodes. The $B^s_{RX,I}$ values of b can be calculated from its own $B^s_{RX,C}$ values and those of its neighbor nodes (again, in general the $B^s_{RX,C}$ values of all nodes in $CN_{\leq d-1}(b)$ would be needed). The $B^s_{RX,C}$ values can in turn be determined by each node from the TX^s values of its neighbors. This way, a can derive $F^s_{TX}(a,b)$ by aggregating the predicates of its neighbors.

V. Assessment of Existing Reservation Protocols

In this section, we apply the local reservation criterion with communication neighborhood to assess existing reservation protocols for wireless networks. In many

protocols, available bandwidth is modeled as an abstract number (*statistical* reservations), for example *Ticket-Based Probing* [3], the *Liao01 Protocol* [4] or *Trigger-Based Distributed Routing* [5]. By their nature, these reservation protocols do not guarantee collision freedom, since bandwidth cannot be reserved exclusively. Therefore, we restrict ourselves to protocols with TDMA approaches, i. e., protocols supporting the exclusive reservation of time slots (*deterministic* reservations).

Some deterministic approaches, namely *Bandwidth Routing* [6], *On-Demand QoS Routing* [7] and *On-Demand Link-State Multi-Path QoS Routing* [8] use CDMA in addition to TDMA to resolve conflicts between neighboring nodes. However, they do not distinguish whether slots are free for sending or free for receiving, which can lead to an unnecessary blocking of slots. Other protocols can be classified as pure TDMA approaches, which in addition make this distinction. The *Forward Algorithm* [9] is based on AODV [10] and calculates local maxima for adjacent links, which are propagated during route discovery. The slot reservation is done during route reply. In the *Liao02 Protocol* [11], each node keeps track of the slots of all nodes in its 2-hop-neighborhood and the corresponding slot states (*reserved* or *free*) in send and receive tables. Information about the 1-hop and 2-hop neighborhood of a node is recorded in a separate table. The reservation is done during route reply. We observe that these reservation protocols consider the slot states $B^s_{RX,I}$, $B^s_{TX,I}$ and $NA^s_{TX,I}$ only for nodes in communication range, but not in interference range. This clearly limits their scope and functionality, as collision freedom cannot be guaranteed despite reservation.

In the following, we will look at two protocols supporting deterministic reservations in more detail, namely the *Race-Free Bandwidth Reservation Protocol* [12] and the *Distributed Slots Reservation Protocol* (DSRP) [13].

A. Race-Free Bandwidth Reservation Protocol

The *Race-Free Bandwidth Reservation Protocol* [12] is an improvement of [11]. It is an on-demand, source-based protocol, whose objectives are to support parallel reservations and to avoid reservation races, which can occur if reservations are processed simultaneously.

The protocol structures time into TDMA frames consisting of control phase and data phase. In the control phase, each node has an exclusive control slot, which can be used to dynamically reserve data slots in the data phase.

For each node in 1-hop and 2-hop neighborhood, send and receive tables recording slot states are maintained. Besides the states *free* and *reserved* of [11], an additional state *allocated* is used for unconfirmed reservations. We note that the distinction between confirmed and unconfirmed reservations is useful for a (distributed) reservation algorithm, but not required in our analysis of the interference problem.

A wait-before-reject strategy is used, which means that a QoS request is not rejected, if enough slots are expected to become available with a predetermined acceptable delay. To realize this, time-to-live timers are used, which reset a slot status from *allocated* to *free* if the corresponding request is not confirmed within a predefined amount of time. Thus, reservation races can occur, if the slot status is set back to *free* too soon.

All nodes periodically broadcast their send and receive tables to their 1-hop and 2-hop-neighbors. In addition, status updates are sent asynchronously when a slot state is changed from *free* to *allocated* or from *allocated* to *reserved*. It follows that the status information of the predicates RX^s and $NA^s_{TX,TX}$ is available, typically with some delay. In addition, the values of the predicates $B^s_{TX,I}$ and $NA^s_{TX,I}$ are available, however, restricted to nodes in communication range. From this, it follows that collision freedom of data frames cannot be guaranteed despite reservations. Also, propagation of slot status information can only take place in the (exclusive) control slot of a node, leading to some delay. Therefore, a QoS request could be started by a neighbor before the status update has been made. This could in fact lead to interference due to double reservations of slots by neighboring nodes, which the protocol claims to eliminate.

B. Distributed Slots Reservation Protocol

The *Distributed Slots Reservation Protocol* (DSRP) [13] is an on-demand slot reservation protocol for QoS routing in TDMA networks. The main objective is the reuse of time slots. For example, slots with least conflict to other mobile hosts or slots used by other mobile hosts can be preferred.

As in [12], time is structured into TDMA frames consisting of control and data subframe, which are subdivided into slots. However, control slots are not exclusively reserved for a particular node, which means that there is contention for medium access, which may cause collisions and unpredictable delays (e. g., of QoS requests).

Besides the hidden and exposed terminal problems, two main problems considered are *slot shortage for self-route* (because of an inappropriate slot choice, a pending QoS request cannot be granted) and *slot shortage for neighboring routes* (because of an inappropriate slot choice, another QoS request cannot be granted).

The slot states forming the (global / local) reservation criterion are considered by the *slot inhibited policies*, but interference is only considered between nodes in communication neighborhood. Since the information is not maintained proactively as in [12], it has to be exchanged when needed. A potential sender x collects information from its neighbors, determines its valid sending slots and forwards them to the potential receiver y, which derives the valid slots for a transmission on the corresponding link. This means that the slot states RX^s, $B^s_{TX,I}$ (limited to nodes in communication range) and $NA^s_{TX,TX}$ are determined by x, and $NA^s_{TX,I}$ (also limited to nodes in communication range) is evaluated by y.

Since this information exchange must happen in the control subframe, there may be collisions and inaccurate information due to delays. Furthermore, since the sending

 ISBN: 978-1-61208-203-5

slots are not reserved in x when they are forwarded to y, another QoS request could allocate these slots. Slot reservation is not done until the QoS reply is propagated through the network.

A *slot adjustment protocol* is used to solve the resulting problems. But since this protocol is only invoked if reservations collide directly at a node, problems could arise if a neighboring node has reserved this slot in the meantime, since this information is only kept in the reserving node.

To alleviate the slot shortage problems mentioned above, *slot decision policies* are used to determine the slots allowing the greatest reuse among the available slots. According to the slot decision policies, the slots used for a certain link are determined by the node three hops apart, in order to find the most suitable slots. But this means that the information on valid slots can already be outdated when the slots to be assigned are calculated.

VI. Related Work

The interference problem in wireless networks has been extensively studied in previous work, using different models to identify conditions for (non-)interference. For a comprehensive survey of interference models in wireless ad-hoc networks, we refer the reader to [14]. The purpose of an interference model is to determine whether a transmission between a pair of nodes may be successful. In general, this depends on many factors, such as spatial placement of nodes, environmental conditions, transceiver and channel characteristics, signal characteristics and propagation, and temporal channel usage. This makes the accurate treatment of interference a complex task. In this section, we focus on network models, which can be classified as graph-based models, physical models, and statistical models:

- *Graph-based models* define a network as a set of vertices representing, e. g., nodes, connected by edges representing, e. g., communication links. The network model we use in this paper belongs to this category.
- *Physical models* capture the characteristics of transceiver and channel with different degrees of detail, taking, e. g., Signal to Interference plus Noise Ratio (SINR) into account.
- *Statistical models* express relevant aspects, e. g., the transmission characteristics, in terms of a probability density function.

Graph-based models may be seen as an abstraction of physical and statistical models, leaving out parameters such as received signal strength or the probability of a successful reception. Different kinds of graph-based models have been applied to interference modeling:

- In the *connectivity graph*, vertices and edges represent nodes and communication links, respectively. In simple interference models, a transmission from a to b is only disturbed by the overlapping transmission of another node $c \neq a$ directly connected to b. In more sophisticated models, the transmission is also disturbed by the overlapping transmission of other nodes with a distance to b of up to 2 or 3 hops.
- In the *interference graph*, vertices and edges often represent nodes and interference links, i. e., interference between nodes. In [15], vertices represent links, and edges model interference between links. In both cases, it is straightforward to identify conditions for (non-) interference.
- In [16], the connectivity and interference graph are merged and augmented by sensing links. Here, a connectivity link implies an interference link, which in turn implies a sensing link. Connectivity links are directed (whereas interference and sensing links are undirected) and exist only if there is at least one transmission according to the link.

In this paper, we have merged the connectivity and interference graph to define the global reservation criterion (Section III) and the local reservation criterion with interference neighborhood (Section IV-A). We have then reduced this graph to a connectivity graph in which interference occurs if a receiver is in maximal 2-hop neighborhood of another sender. This way, it is feasible to define a local reservation criterion that is actually implementable (Section IV-B). We have decided against modeling sensing links, as concurrent transmissions of nodes in sensing range, but not in interference range would not be harmful.

VII. Conclusion and Future Work

In this paper, we have analyzed the interference problem in wireless networks, considering exclusive reservation schemes based on TDMA, where time slots are synchronized network-wide with an upper bound for clock offset. For this analysis, we have used a graph-based network model with edges representing communication and interference links. In a first step, we have defined an obvious global reservation criterion that solves the interference problem, however, at the expense of global up-to-date topology and reservation status. In a second step, we have rewritten the global criterion into an equivalent local form, thereby reducing the status information to solve the interference problem. In a third step, we have rewritten the local form, assuming that the interference range is limited to maximal 2-hop communication neighborhood, and have argued that this local form is actually implementable. Based on this local form, we have assessed a selection of existing reservation protocols and have identified a number of shortcomings.

In our future work, we will broaden our study of existing reservation protocols and develop a taxonomy for their assessment. Furthermore, we will make an effort to develop reservation protocols that solve the interference problem by implementing the second local reservation criterion. We note that it is not clear which additional assumptions are to be made to solve the problem of inaccurate or outdated topology and reservation status.

References

[1] J. Schiller, *Mobile Communications*, 2nd ed. Boston: Addison-Wesley, May 2003.

 ISBN: 978-1-61208-203-5

[2] R. Gotzhein and T. Kuhn, “Black Burst Synchronization (BBS) - A Protocol for Deterministic Tick and Time Synchronization in Wireless Networks,” *Computer Networks*, vol. 55, no. 13, pp. 3015–3031, 2011.

[3] S. Chen and K. Nahrstedt, “Distributed Quality-of-Service Routing in Ad-Hoc Networks,” *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 8, pp. 1488–1505, August 1999.

[4] W.-H. Liao, Y.-C. Tseng, S.-L. Wang, and J.-P. Sheu, “A Multi-path QoS Routing Protocol in a Wireless Mobile Ad Hoc Network,” in *ICN ’01: Proceedings of the First International Conference on Networking-Part 2*. London, UK: Springer-Verlag, 2001, pp. 158–167.

[5] S. De, S. K. Das, H. Wu, and C. Qiao, “Trigger-Based Distributed QoS Routing in Mobile Ad Hoc Networks,” *Mobile Computing and Communications Review*, vol. 6, no. 3, pp. 22–35, 2002.

[6] C. R. Lin and J.-S. Liu, “QoS Routing in Ad Hoc Wireless Networks,” *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 8, pp. 1426–1438, August 1999.

[7] C. R. Lin, “On-Demand QoS Routing in Multihop Mobile Networks,” in *In Proc. IEEE Infocom*, 2001, pp. 1735–1744.

[8] Y.-S. Chen, Y.-C. Tseng, J.-P. Sheu, and P.-H. Kuo, “An On-Demand, Link-State, Multi-Path QoS Routing in a Wireless Mobile Ad-Hoc Network,” *Computer Communications*, vol. 27, no. 1, pp. 27–40, 2004.

[9] C. Zhu and M. S. Corson, “QoS Routing for Mobile Ad Hoc Networks,” in *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2. IEEE, 2002, pp. 958–967.

[10] C. E. Perkins and E. M. Royer, “Ad-hoc On-Demand Distance Vector Routing,” in *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applicatons*, New Orleans, LA, Feb. 1999, pp. 90–100.

[11] W.-H. Liao, Y.-C. Tseng, and K.-P. Shih, “A TDMA-based Bandwidth Reservation Protocol for QoS Routing in a Wireless Mobile Ad Hoc Network,” in *Proceedings of IEEE ICC*, April–May 2002.

[12] I. Jawhar and J. Wu, “A Race-Free Bandwidth Reservation Protocol for QoS Routing in Mobile Ad Hoc Networks,” in *HICSS*, 2004.

[13] K.-P. Shih, C.-Y. Chang, Y.-D. Chen, and T.-H. Chuang, “Dynamic bandwidth allocation for QoS routing on TDMA-based mobile ad hoc networks,” *Computer Communications*, vol. 29, no. 9, pp. 1316–1329, 2006.

[14] P. Cardieri, “Modeling Interference in Wireless Ad Hoc Networks,” *IEEE Communications Surveys and Tutorials*, vol. 12, no. 4, pp. 551–572, 2010.

[15] R. Gupta, Z. Jia, T. Tung, and J. Walrand, “Interference-aware QoS Routing (IQRouting) for Ad-Hoc networks,” in *IN PROC. ICC 2005, SEOUL, KOREA*, 2005, pp. 1–6.

[16] J. L. Sobrinho and A. S. Krishnakumar, “Quality-of-Service in Ad Hoc Carrier Sense Multiple Access Wireless Networks,” *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 8, pp. 1353–1368, August 1999. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=779919 [retrieved: April, 2012]

 ISBN: 978-1-61208-203-5

QoS-aware Resource Allocation for In-band Relaying in LTE-Advanced

Thiago Martins de Moraes, Arturo Antonio Gonzalez, Muhammad Danish Nisar, Eiko Seidel
Nomor Research GmbH
Brecherspitzstr. 8, 81541 Munich, Germany
Email: {moraes,gonzalez,nisar,seidel}@nomor.de

***Abstract*—Fulfilling the heterogeneous quality of service (QoS) requirements of individual users is a central theme in future wireless networks. The addition of relay nodes introduces some new challenges towards achieving this target. In this work, we study the problem of resource allocation in advanced-relay scenarios. To this end, we propose the design of an efficient QoS-aware scheduler that strikes a balance between the latency and the bit rate requirements of individual traffic flows. The proposed scheduler is implemented at the donor eNB and the relay node by adapting to the additional challenges introduced by the relay node's wireless backhaul link. Finally, via system level simulations for the downstream direction emulating traffic with different bit rate and latency requirements, we demonstrate that our algorithm is able to multiplex mixed traffic with small or no violation of the individual QoS requirements, thereby achieving significant gains over baseline approaches.**

***Index Terms*—Quality of service; QoS; Relay; LTE; Resource allocation; Scheduler; Backhaul link; VoIP; Video; Delay budget; In-band Relays.**

I. INTRODUCTION

One of the key functionalities to improve the cell-edge user throughput in the Long Term Evolution Advanced (LTE-A) is the support of relay nodes (RN), [1], which are low power transmitting nodes typically deployed at the cell-edge or coverage holes. Currently, LTE-A supports only non-transparent relays, or type-1 relays. These type-1 RNs appear to the User Equipment (UE) as an eNB with an independent coverage area whereas to the evolved NodeB (eNB), or base station, they appear as an UE with special capabilities.

In contrast to an eNB, the RNs do not have a wired connection to the core network, hence all data from/to the relay has to be forwarded through a wireless backhaul link established between the RN and a regular eNB, which, in this scenario is called Donor eNB (DeNB). Non-transparent (type-1) relays are classified depending on the way the backhaul and access links are multiplexed: When the backhaul and the relay access radio link use the same frequency band and they are segregated in the time domain, the relay is referred to as in-band relay. The relays are classified as out-band, when the backhaul link and the relay access links are allocated to different frequency bands. For in-band relays, special frames are reserved for the RN communication with the DeNB, the so-called MBSFN frames. During these frames, the RNs are only allowed to receive data from the DeNB, and they refrain from transmitting in order to avoid self interference.

In an advanced relay scenario, the eNBs and RNs have to host a mix of different traffic types. For instance, while some users are using Voice over IP (VoIP) services, some other are browsing, others have active FTP services or have active video streaming services. Each of these services has different requirements in terms of bit rate, latency, jitter, delay, etc. The presence of heterogeneous quality of service (QoS) requirements, call for the design of sophisticated resource allocation algorithms, that fulfill individual QoS requirements and ensure that they are not violated.

Although all architectural aspects are defined by the standard, the resource allocation is still an interesting topic for research. The fundamental questions that have to be answered by a resource allocation algorithm, especially in a relay enhanced scenario are:

- How to split/partition the resources at the DeNB between the macro UEs (M-UEs) and the wireless backhaul link?
- How to allocate resources at the DeNB and at RN, in order to satisfy the QoS constraints of the R-UEs accumulated over both hops?
- How to coordinate the resource allocation at the DeNB and the RN, so as to reduce the effect of interference on the access links?

In this work, via the design of our novel resource allocation algorithm, we focus primarily in answering the first two of these aspects. The last one is being considered for future work.

A. Related Work

Resource allocation for satisfaction of heterogeneous QoS requirements in presence/absence of relays has been an active area of research over the past few years. In [2], Liu *et. al.* propose a manner of scaling the delay influence on a QoS scheduler in a multi-hop wireless mesh network. However, this work does not consider the relay deployments in LTE-A, and its specific requirements.

The support of QoS in LTE-Advanced for mixed traffic is studied in [3] which divides the traffic types into two types: real-time traffic and non-real-time traffic. The flows belonging to the non-real-time traffic are scheduled based on a proportional fair metric which is scaled by the respective QoS requirements, and a scaled Max C/I approach is used to scheduled the real-time traffic. But this work also does not consider the resource allocation challenges in relay enhanced scenarios.

 ISBN: 978-1-61208-203-5

In [4] and [5], the issue of resource allocation for in-band relays was studied. However, both of these contributions focus on resource allocation only for flows with no QoS requirements.

B. Our Contribution

We focus on the problem of QoS-aware resource allocation in relay-aided future wireless networks. The description in this paper is oriented towards in-band relaying in LTE-A, however the proposed resource allocation mechanism is generic enough, and can be applied to other relay enhanced network deployments as well, with heterogeneous QoS requirements.

The remainder of this paper is organized as follows: In Section II, while highlighting the challenges associated with QoS-aware resource allocation in an enhanced relay scenario, we provide the description of the proposed algorithm. In Section III, a short description of the simulation setup is given, followed by the presentation and discussion of the main simulations results, including comparison with baseline approaches, in Section IV. Finally, a summary of the major achievements and outlook for future work in this area are provided.

II. Proposed QoS-aware Resource Allocation in Relaying Scenarios

A. Two-stage Scheduler Structure

We propose to employ a two-stage scheduler for the OFDM based downlink LTE system. Our LTE-A scheduler is therefore composed of two main stages: the time-domain (TD) stage and the frequency domain (FD) stage. The main function of the TD stage scheduler is set up a candidate list of users which are to be scheduled in a particular scheduling round. To this end, the users are sorted according to predefined rules (e.g., round-robin, proportional fair) or a metric computations based on their individual requirements. The TD scheduler also serves another propose: it reduces the number of users that will be forwarded to the next phase (the FD stage) thereby reducing the complexity of the resource allocation process. After the users are sorted, the created list is then forwarded to the FD stage scheduler. It is worth-mentioning here that we assign pending re-transmissions a higher priority so that they are always placed on the top of the TD candidate list.

The FD stage scheduler is responsible for the actual allocation of the frequency resource blocks to the users. All the resources are visited one by one, and the user with the highest metric is allocated the current resource block. After each allocation, the residual QoS requirements (urgency) of the scheduled user is updated. The allocation process ends when all the resources have been allocated, or no data is available.

B. QoS Requirements

The QoS requirements are specified in LTE-A typically via predefined Quality Class Indicator (QCI) [6]. QCI is an scalar which refers to a set of fix service parameters that are used in the packet forwarding decisions. Each one of the nine possible QCI values defines the

- delay budget: the maximum acceptable packet delay,
- maximum block error rate,
- service priority index: a scalar ranging from one to nine; the higher the index, the lower is the service priority.
- some QCIs also specify Guaranteed Bit-rate (GBR); the specification of GBR for a particular service is left open for the service provider.

In the following, we consider three of these QoS requirements namely, the delay budget, the GBR, and the priority index, while designing our QoS aware resource allocation algorithm.

C. Proposed Scheduling Metric

We start our studies by defining the QoS-scheduling metric which is used to sort the UEs according to the data urgency. The first step towards the definition of our scheduling metric is to define the delay coefficient $\omega_{\mathrm{d}}(n,t)$ which represents the effect of the packet delay on the metric computation: For a packet belonging to the flow n, we define the delay coefficient at the time instant t, as follows

$$\omega_{\mathrm{d}}(n,t) = \exp\left(\beta \frac{\mathrm{d}_{\mathrm{HOL}}(n,t)}{D_{\mathrm{Profile}}(n)}\right) \tag{1}$$

where $d_{\mathrm{HOL}}(n,t)$ is the head of line (HOL) delay of the flow n at time t, $D_{\mathrm{Profile}}(n)$ is the delay requirement of the flow n as specified in its QCI, and β is an scalar factor which pronounces the effect of the exponential function.

Using the delay metric from (1), we now define the overall QoS-scheduling metric as:

$$m_{\mathrm{QoS}}(n,t) = \max\left[\left(\frac{\mathrm{GBR}(n)}{\bar{\mathrm{R}}(n,t)}\right),1\right]^{\rho} \cdot \frac{\omega_{\mathrm{d}}(n,t)}{\mathrm{P}(n)} \tag{2}$$

where $\bar{R}(n,t)$ is the average throughput of the flow n over past few intervals, while $\mathrm{GBR}(n)$ and $P(n)$ are respectively the guaranteed bit rate and the service priority index of the flow n as specified in its QCI. ρ is a factor which emphasizes the rate metric if their $\bar{R}(n,t)$ is lower then the required $\mathrm{GBR}(n)$.

We observe that the proposed scheduling metric in (2) consists of two main factors.

- The rate factor defined by the term inside max function forces the fulfillment of the GBR(n): While $\bar{R}(n,t)$ is smaller than $\mathrm{GBR}(n)$, the factor increases the metric, otherwise it has no effect on the metric.
- The second factor is $\omega_{\mathrm{d}}(n,t)$ which has small influence in the metric, if the delay is low, but increases the metric exponentially as the delay gets closer to the packet deadline.

We propose to employ the scheduling metric in (2) for sorting the users in the TD scheduling stage and also to assign resources in the FD scheduling stage, as discussed in Section II-A. For our simulation results in the next sections, we choose $\beta = 5$ and $\rho = 4$, respectively to pronounce the exponential effect of delay budget and GBR constraints.

 ISBN: 978-1-61208-203-5

D. Addressing Additional Challenges in Relaying Scenarios

The introduction of the second hop in a relay enhanced network introduces additional challenges for the support of QoS-aware services.

First and foremost, in such networks, the downlink resource allocation for the relay user equipments (R-UEs) has to be performed in two stages: In the first stage, while serving the M-UEs, the DeNB transfers the user data from its buffers to the serving relay node by scheduling resources to the backhaul link. Afterwards, each RN schedules the received data to their subordinate R-UEs. In other words, a packet destined to a R-UE has to undergo two scheduling process, and each of them results in extra packet delay. Therefore, a special attention has to be given to the flows that are multiplexed through the RNs. Hence, we define for the R-UEs the accumulated QoS requirements as follows:

$$T_{\text{DeNB-RN}} + T_{\text{RN-UE}} \leq D_{\text{Profile}}(n) \tag{3}$$

$$\frac{B_{\text{UE}}}{T_{\text{DeNB-RN}} + T_{\text{RN-UE}}} \geq \text{GBR}(n) \tag{4}$$

where $T_{\text{DeNB-RN}}$ is time interval between the packet arrival at the DeNB until it is received by the RN, $T_{\text{RN-UE}}$ is the interval between the reception in the RN and the time that the packet is received at the R-UE, and $D_{\text{Profile}}(n)$ is the delay requirement of the flow n. Moreover, B_{UE} is the volume of data transferred in the time interval $T_{\text{DeNB-RN}} + T_{\text{RN-UE}}$ and $\text{GBR}(n)$ is the rate requirement of the used flow.

Secondly, for in-band relays, the resources in the backhaul link must be scheduled only during the MBSFN frames. During these frames, the DeNB has to decide on how to partition the available resources between the wireless backhaul link and to the M-UEs. In this regard, we propose to bundle all the relay UEs (R-UEs) with the same QCI (similar QoS requirements) into a single flow with aggregate service requirements. Afterwards, the scheduler depending on the urgency of the M-UEs and the aggregated R-UEs flows decides whether a resource block is to be given to the backhaul link or to the access link.

As mentioned above, from the scheduling perspective of the DeNB, the backhaul link is a normal UE link, with QoS requirements of all underlying flows merged into single/multiple backhaul link "super flow". Based on (3) and (4), we define the GBR requirements for the backhaul link as:

$$\text{GBR}(B) \geq \sum \text{GBR}(n), \tag{5}$$

i.e., a sum of GBRs of all underlying flows. Furthermore, the delay requirement at each scheduling node for the virtual "super flow" is defined as:

$$D_{\text{QoS-B}} = \frac{D_{\text{Profile}}}{N}, \tag{6}$$

i.e., we split the QCI-specified delay budget equally among the N hops. Thus, we force the scheduler to send the packet earlier than what it would normally consider in a single hop scenario, thereby allowing the second hop scheduler the possibility of delivering the data before the packet deadline. In our scenario, a packet has to be forwarded through two hops, namely backhaul and RN access links, thus the maximum delay allowed at the first hop (DeNB) scheduler is restricted to be:

$$D_{\text{QoS-DeNB}} = \frac{D_{\text{Profile}}}{2}. \tag{7}$$

Note that, at the second hop (RN) scheduler, we extract the information of the delay that the packet has already went through, and adjust the packet HOL delay timer such that the total packet delay across both hops is still within the D_{Profile} as specified in the service QCI.

III. Simulation Setup

A. Deployment Scenario

The deployment scenario we have considered for our performance evaluations is a single macro-cell with 1 DeNB and 1, 2, or 4 relay nodes attached to it. Note that in the remainder of this work, we will only show results for the case of 2 relay nodes due to the limited space. The macro-cell is modeled as a hexagon with the DeNB at one corner and the RNs located along the opposite side. The hexagonal shape thus resembles one of the three sectors served by the DeNB, which corresponds to the reduced single-cell layout specified in [7]. The interference from all neighboring cells at non-negligible distance to the macro-cell is also considered in the model.

Within the macro-cell, a so-called "hot-spot" scenario is assumed: A total of 25 UEs are placed such that a pair of 2 UEs (the *R-UEs*) always ends up in the coverage region of a relay node and the remaining 21 UEs (the *M-UEs*) in the coverage region of the DeNB. In addition, all UEs are periodically relocated randomly within their respective coverage regions. Due to this automatic re-positioning, no explicit motion model or handover procedure has been considered in this work.

Finally, we would like to emphasize that the same sequence of UE positions is replayed when simulating the macro-cell without any relay nodes, such that the performance can be directly compared.

B. Macro-cell Configuration

The used channel model complies with the 3GPP *Case-1* for urban macro-cells with an inter-site distance (ISD) of 500 m, as specified in [1], Table A2.1.1-1. The respective configuration parameters for the DeNB-UE link are contained in this table and in the following table, Table A2.1.1-2. We will only recall the most important ones here for the sake of completeness:

- Carrier frequency: 2 GHz
- Duplexing and bandwidth: FDD with 10+10 MHz
- TTI duration: 1 ms
- Speed: 3 km/h
- Penetration loss: 20 dB
- Path loss: Only NLOS term (for macro to UE) used
- 3-D antenna pattern with 15 degree electrical downtilt

 ISBN: 978-1-61208-203-5

- DeNB antenna height: 32 m
- UE antenna height: 1.5 m
- Minimum distance between DeNB and UE: 35 m
- DeNB Tx power: 43 dBm

While fast fading is considered in our model, we have omitted the log-normal shadowing for now, since the usual 1-D correlated model proposed in [1] does not lead to meaningful results when applied across a 2-D plane.

C. Relay Configuration

We also assume 3GPP *Case-1* here with 2 outdoor relays, thus complying with [1], Table A2.1.1.2-2. The respective configuration parameters for the RN-UE link are contained in Table A2.1.1.2-3 and A2.1.1.4-3. We will also recall only the most important ones here:

- Carrier frequency, duplexing, bandwidth, TTI duration, speed, and penetration loss: same as for macro-cell
- Path loss: Only NLOS term (for relay to UE) used
- 2-D omni-directional antenna pattern with 5 dBi gain, 2Tx and 2Rx antenna ports
- RN antenna height: 5 m
- UE antenna height: same as for macro-cell
- Minimum distance between RN and UE: 10 m
- RN Tx power: 30 dBm

Note that for the DeNB-RN link, we assume a gain of 5 dB to account for the quasi-stationary reception conditions. A typical signal to interference and noise ratio (SINR) map of the cell, indicating also the RN locations, is shown in Figure 1.

D. Traffic Model

In our evaluation, we employ two different traffic types with their distinct QoS requirements:

- VoIP traffic are emulated using a Constant Bit-rate (CBR) traffic generator which creates traffic at rate of 128 kbps. The QCI for this type of traffic is defined as QCI-1 [6]: The packets are allowed to have a maximum end-to-end delay of 100 ms, and the service priority is defined as 2.
- A second group of users are using video streaming services. Due to the limitations of the mobile devices, such as processing capabilities, we have chosen to limit this service to a rate of 256 kbps. The QCI for this type of traffic is defined as QCI-3 [6]: Maximum packet end-to-end delay is 300 ms, and the service priority is 5.

The desired CBR is created by our traffic generator using a fixed packet size of 256 bytes at appropriate time intervals: 8 ms for video and 16 ms for VoIP traffic.

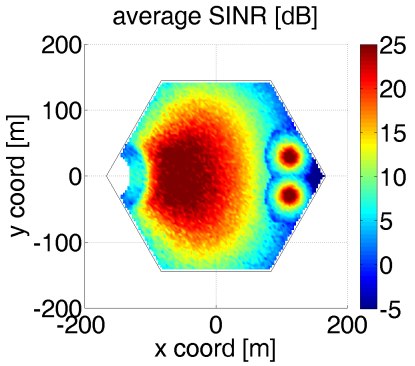


Fig. 1. Typical SINR map of DeNB plus 2 relay nodes (DeNB+2RN) scenario. Two dark red spots towards the far end of hexagon indicate the locations of the two relay nodes.

IV. Performance Results

A. Baseline macro-cell-only scenario

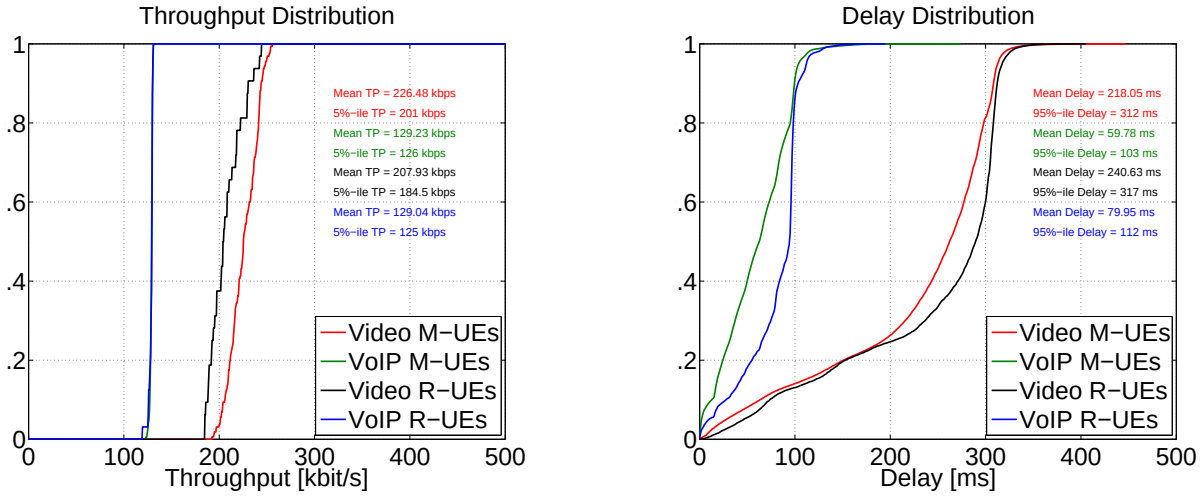
We start our analysis by comparing the behavior of different flows for the case where DeNB is the only transmitting node serving the macro-cell. In order to provide a better understanding of the system behavior, we have divided the users into two groups:

- One group is formed by the users that would be in the coverage area of the RNs had they been activated, and we label these as R-UEs.
- The second group consists of the macro users (M-UEs) which always connect to the DeNB regardless of the simulated scenario (DeNB-only, or DeNB+2RNs).

Figure 2 depicts the Cumulative Distribution Function (CDF) of the per-user throughput and delay for the two traffic types (VoIP and Video) in each user category (M-UEs and R-UEs) for the DeNB-only scenario with the proposed QoS-aware resource allocation strategy. From the throughput CDFs in Figure 2(a), we observe that VoIP users are able to achieve their designated GBR requirement to a large extent. The mean and 5%-ile TP values are 129 kbps and ca. 125 kbps for both M-UEs and R-UEs. However, the achieved throughput of the video UEs is significantly less as compared to the configured GBR. In contrast to the configured 256 kbps for video users, the mean and 5%-ile TP values for M-UEs are 226 kbps and 201 kbps respectively, while for R-UEs these are only around 208 kbps and 184 kbps respectively. We observe that R-UEs suffer more as compared to the M-UEs, and this owes to the fact that they are served over poorer radio conditions. The impact on M-UEs is basically a consequence of R-UEs throttling the performance of the overall system.

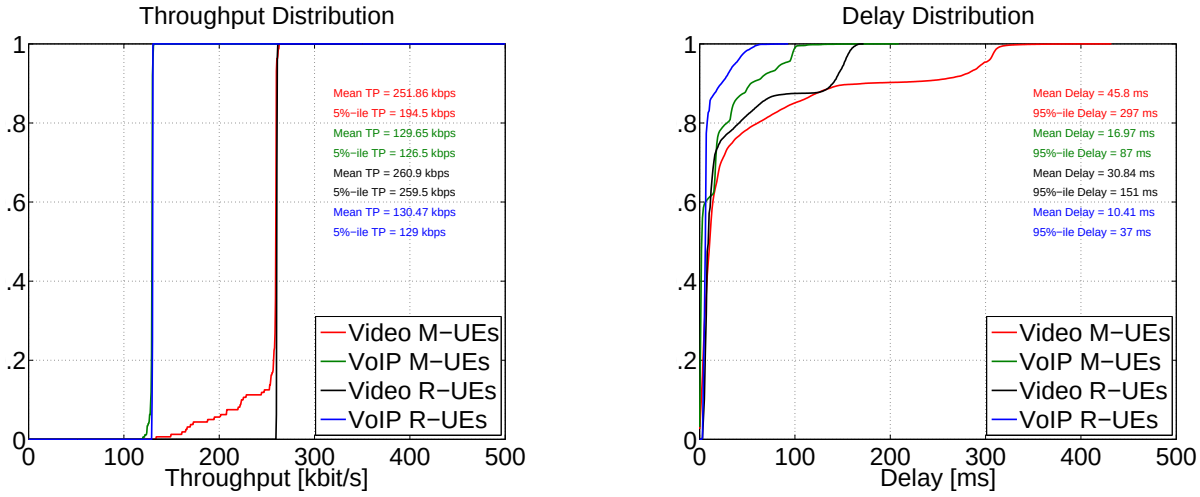
For the same scenario, we now focus on the end-to-end packet delay CDFs in Figure 2(b). In line with our observations regarding Figure 2(a), we note that the packet delays for both traffic types and for both UE groups is quite high, and quite often approaching the ultimate delay budget deadline, after which the packets are discarded at the scheduler. The VoIP 95%-ile delays being 103 ms and 112 ms respectively for M-UEs and R-UEs. Similarly, the video 95%-ile delays are around 312 ms and 317 ms respectively for M-UEs and R-UEs indicating that there is non-negligible network congestion and packet discard ongoing in the system. Note that the slight overshoot of delays beyond the delay budget can be attributed to the extra delay caused by re-transmissions.

 ISBN: 978-1-61208-203-5



(a) Throughput CDFs (b) Delay CDFs

Fig. 2. DeNB-only scenario with proposed QoS-aware resource allocation. Throughput and delay CDF comparisons for different UE groups (M-UEs and R-UEs) and traffic types (VoIP and Video).



(a) Throughput CDFs (b) Delay CDFs

Fig. 3. DeNB+2RN scenario with proposed QoS-aware resource allocation. Throughput and delay CDF comparisons for different UE groups (M-UEs and R-UEs) and traffic types (VoIP and Video).

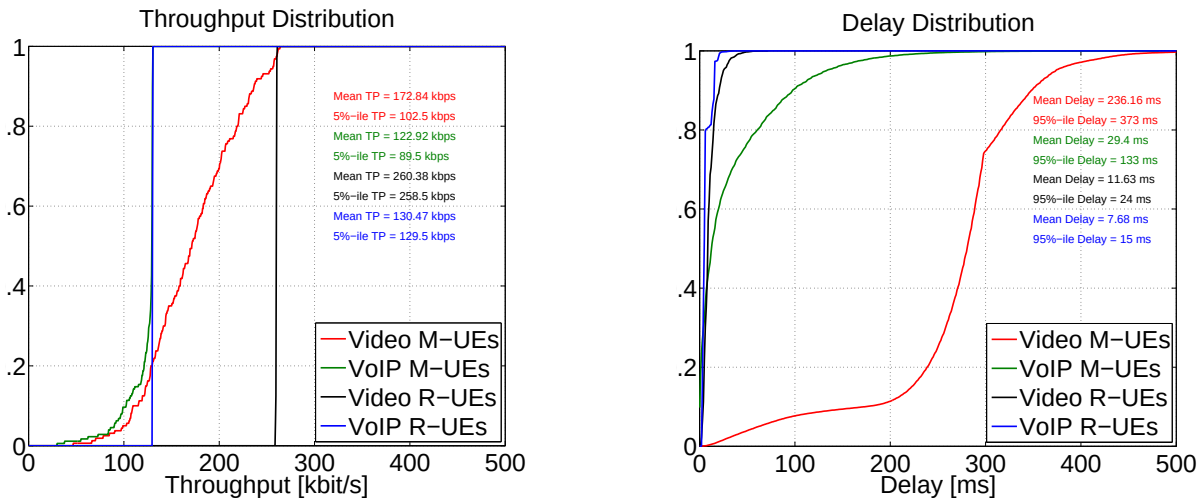
B. Relay enhanced scenario (DeNB+2RN) with proposed QoS-aware resource allocation

From results in last sub-section, we conclude that though the QoS-aware resource allocation attempts to fulfill the QoS requirements, the level of satisfied users is quite low owing to the congestion in network. In this sub-section, we consider the same traffic pattern, but now in a relay enhanced scenario, where two relay nodes are deployed to assist the DeNB.

In Figure 3, we present the throughput and delay CDFs for the DeNB+2RN scenario with the proposed QoS-aware resource allocation strategy. In contrast to Figure 2(a) for the TP CDF in DeNB-only scenario, in Figure 3(a), we observe that the rate requirements of both traffic types are fulfilled to a large extent for both UE categories. The mean and 5%-ile TP values for VoIP traffic type are 130 kbps and ca. 128 kbps for both M-UEs and R-UEs. For the video UEs, we observe that performance of R-UEs improve considerably as compared to that in Figure 2(a). The mean and 5%-ile TP values being 261 kbps and 260 kbps respectively. For the video M-UEs, we observe that there is still a fraction of M-UEs that are not able to achieve the designated GBR. The mean and 5%-ile TP values for video M-UEs are 252 kbps and 195 kbps respectively.

Next, in Figure 3(b), we plot the CDF of the end-to-end packet delays for proposed scheme in DeNB+2RN scenario. Note that the end-to-end packet delay in this scenario corresponds to a sum of packet delays experienced over both hops. We observe that adding relay nodes helps the system

 ISBN: 978-1-61208-203-5



(a) Throughput CDFs

(b) Delay CDFs

Fig. 4. DeNB+2RN scenario with conventional resource allocation (static resource partitioning plus proportional fair scheduler). Throughput and delay CDF comparisons for different UE groups (M-UEs and R-UEs) and traffic types (VoIP and Video).

significantly, in comparison to the DeNB-only case in Figure 2(b). The mean packet delays for both services are significantly reduced. The reduction for R-UEs delay comes from the fact that the RN is able to serve them over better radio conditions, while the reduction for M-UEs delay comes as a consequence of reduction in resource consumption for R-UEs. Besides the significant reduction of mean packet delays, 95%-ile delays are also reduced. The VoIP 95%-ile delays are reduced to 87 ms and 37 ms respectively for M-UEs and R-UEs. Similarly, the video 95%-ile delays reduce to around 297 ms and 151 ms respectively for M-UEs and R-UEs. The video R-UEs having 95%-ile delay of around 151 ms owes to the fact that the major fraction in end-to-end delay for R-UEs comes from the first hop, which in this case is assigned a delay budget deadline of 150 ms. Similar remark holds for the VoIP R-UEs.

In summary, we observe from Figure 3 that once sufficient radio resources are available, the proposed QoS-aware resource allocation mechanism facilitates to a very large extent the satisfaction of heterogeneous QoS requirements in terms of both latency and bit rate simultaneously.

C. Relay enhanced scenario (DeNB+2RN) with conventional resource allocation

Finally, in this sub-section, we present the throughput and delay CDFs for the same relay enhanced scenario (DeNB+2RN), as in last sub-section, but with conventional resource allocation in place of QoS-aware resource allocation. To this end, we pursue static resource partitioning [4] at the DeNB between the macro-access link and the backhaul link. This static resource partitioning incorporates the fact that the video users have twice the rate requirement of the VoIP users. A regular proportional fair scheduler is employed with in the partitioned spectrum to assign resources to various competing users.

From Figure 4, we observe an imbalance with regard to QoS satisfaction of users. In Figure 4(a), we note that though the TP requirements for R-UEs are easily met, the observed mean and 5%-ile values for M-UEs are only around 173 kbps and 103 kbps for video, and around 123 kbps and 90 kbps for VoIP users. Similar observations can be made from Figure 4(b) depicting the delay CDF; we note that though the delays for R-UEs are lower as compared to Figure 3(b), but the degradation in the performance of M-UEs is rather drastic.

D. Comparison between proposed and conventional resource allocation

In order to summarize the effectiveness of the proposed QoS-aware resource allocation scheme for relay enhanced networks, we present in this sub-section a comparison with the conventional scheme, in terms of the fraction of satisfied users. To this end, we define two alternative measures:

- μ_{TP} = Fraction of satisfied users w.r.t. throughput, defined as the fraction of users (or samples) that achieve up to 95% of the configured GBR.
- μ_{DLY} = Fraction of satisfied users w.r.t. delay, defined as the fraction of users (or samples) that are served within the configured delay limits.

For the DeNB+2RN scenario, the fraction of satisfied users w.r.t. TP and delay are presented in Table I for the proposed and conventional schemes. Note that the values are obtained respectively from Figure 3 and Figure 4. It can be seen that especially for M-UEs, the fraction of satisfied users is appreciably increased by employing the proposed QoS-aware resource allocation scheme. For instance, the number of satisfied video UEs increase from 8.1% to 88.8% w.r.t. the achieved throughput, and from 74.7% to 95.3% w.r.t. the experienced packet delay.

 ISBN: 978-1-61208-203-5

TABLE I
DENB+2RN SCENARIO: FRACTION OF SATISFIED USERS μ_{TP} AND μ_{DLY} FOR THE PROPOSED QOS-AWARE (PROP.) VS. THE CONVENTIONAL STATIC RESOURCE PARTITIONING BASED (CONV.) RESOURCE ALLOCATION

UE type	Fraction of satisfied users w.r.t.			
	achieved throughput		experienced delay	
	Conv.	Prop.	Conv.	Prop.
Video M-UEs	8.1%	88.8%	74.7%	95.3%
VoIP M-UEs	80.0%	99.4%	90.5%	99.0%
Video R-UEs	100.0%	100.0%	100.0%	100.0%
VoIP R-UEs	100.0%	100.0%	100.0%	100.0%

V. CONCLUSION AND FUTURE WORK

In this work, we have analyzed the performance of a QoS-aware scheduler for in-band relaying scenario in LTE-Advanced systems. The proposed scheduler addresses two major issues. First, it guides on how to split resources between macro users and the backhaul link at the DeNB scheduler via the concept of "super flows". Secondly, it facilitates the satisfaction of QoS constraints for the R-UEs that are scheduled in two distinct hops. The advantage of the proposed resource allocation scheme in both these aspects has been confirmed via system level simulations, which show significant gains over conventional approaches, in terms of the fraction of satisfied users.

Further work in this area will target an extension of the overall resource allocation framework to interference coordination strategies for relays. We believe that coupling of the proposed scheme with an efficient interference coordination on access links [8] will lead to a better performance of the cell-edge users especially for scenarios with a large number of relay nodes (i.e., 4-10 RNs), where the interference inside the macro-cell can easily degrade the overall system performance.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Commission's seventh framework programme FP7-ICT-2009 under grant agreement n^{o} 2472223 also referred to as ARTIST4G.

REFERENCES

[1] 3GPP TS 36.814, "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects," 3rd Generation Partnership Project (3GPP), Technical specification.

[2] C. H. Liu, A. Gkelias, Y. Hou, and K. K. Leung, "A Distributed Scheduling Algorithm with QoS Provisions in Multi-hop Wireless Mesh Networks," in *4th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2008)*, Avignon, France, Oct 2008, pp. 253 – 258.

[3] R. Kausar, Y. Chen, K. K. Chai, L. Cuthbert, and J. Schormans, "QoS Aware Mixed Traffic Packet Scheduling in OFDMA-based LTE-Advanced Networks," in *The Fourth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies UBICOMM 2010*, Florence, Italy, Oct 2010, pp. 53 – 58.

[4] G. Liebl, T. M. de Moraes, A. Soysal, and E. Seidel, "Fair Resource Allocation for Inband Relaying in LTE-Advanced," in *2011 8th International Workshop on Multi-Carrier Systems & Solutions (MC-SS)*, Herrsching, Germany, May 2011, pp. 1 – 5.

[5] Z. Ma, W. Xiang, H. Long, and W. Wang, "Proportional Fair Resource Allocation for LTE-Advanced Networks with Type I Relay Nodes," in *IEEE International Conference on Communications ICC2011*, Kyoto, Japan, Jun 2011, pp. 1 – 5.

[6] 3GPP TS 23.401, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," 3rd Generation Partnership Project (3GPP), Technical specification.

[7] Artist4G, "D5.1 - Scenarios, KPIs and Evaluation Methodology for Advanced Cellular Systems," Deliverable, Jun 2010.

[8] G. Liebl, T. M. de Moraes, A. Gonzalez, and M. D. Nisar, "Centralized Interference Coordination in Relay-Enhanced Networks," in *IEEE Wireless Communications and Networking Conference (WCNC 2012)*, Paris, France, Apr 2012.

 ISBN: 978-1-61208-203-5

A Bandwidth Reservation Method for IPTV Service

Yong-do Choi
Mobile Communication Engineering
Kyungpook National University
Daegu, Korea
ydchoi@mmlab.knu.ac.kr

Zhi-Bin Yu, Jae-hyun Jun
Electrical Engineering and Computer Science
Kyungpook National University
Daegu, Korea
{zbyu, jhjun}@mmlab.knu.ac.kr

Sung-ho Kim
School of Computer Science and Engineering
Kyungpook National University
Daegu, Korea
shkim@ knu.ac.kr

***Abstract*— A growth of high speed Internet increases network traffic and applications that use voice, data, and multimedia services. Among those, Internet Protocol TeleVision (IPTV) service is rapidly proliferating all over the world. However, network infrastructure cannot accommodate the growth of IPTV. Therefore, this study suggested a bandwidth reservation mechanism using a traditional protocol to achieve a more stable IPTV service. We referenced the Multiple Stream Reservation Protocol (MSRP) that is a bandwidth reservation protocol in IEEE 802.1AVB and implemented a similar reservation mechanism in IPTV service environment; however, but our own mechanism improved the bandwidth reservation fail situations that are not supported by the original MSRP. Therefore, the proposed method is more stable for bandwidth reservation in the IPTV service environment. We examined the proposed method with network simulator, OPNET, and compared an end-to-end delay via the original IPTV service and with the end-to-end delay using our bandwidth reservation.**

Keywords - QoS; bandwidth reservation; IPTV

I. INTRODUCTION

The growth of high speed Internet increases network traffic and applications that use voice, data, and multimedia services. Further, depending on the development of terminal devices capable of playing multimedia - such as Mp3 players, Portable Multimedia Players (PMPs), smart phones, and navigation etc. - many network application services have been created. Among those, Internet Protocol TeleVision (IPTV), which is highly regarded as a killer application service, is rapidly increasing in the world. Sufficient network resources are needed for supporting IPTV services [1]. However, the current network resource is insufficient due to heavy Peer-to-Peer (P2P) traffic and other traffic getting into network; further, there are times when average network service is impossible to achieve due to malicious traffic such as Distributed Denial of Service (DDoS). To achieve a stable IPTV service against harmful traffic, network bandwidth reservation is needed. The current IPTV service offers guaranteed Quality of Service (QoS) by Internet Service Providers (ISPs) to subscribers, but QoS is not guaranteed from subscriber networks to each user because the network transmission policy is best-effort [2].

The Institute of Electrical and Electronics Engineers 802.1 Audio/Video Bridging Task Group (IEEE 802.1 AVB TG) is carrying out research among Ethernet-based digital media devices. First, high quality synchronization services are provided among several digital media devices in LANs. Second, there is a mechanism to make reservation resources for each service in addition to sets of default rules for managing resources. A third kind of research is on a traffic forwarding method through reserved bandwidth [3]. Our study is aimed to achieve Multiple Stream Reservation Protocol (MSRP) which is a kind of bandwidth reservation applied to the IPTV service using Internet Group Management Protocol (IGMP). However, MSRP does not handle some failed bandwidth reservation situations in the IPTV service environment. Therefore, we improved the original MSRP to handle failed bandwidth reservation situations, and our proposed mechanism can support higher QoS than traditional IPTV service.

The remainder of the paper is organized as follows. In Section II, we introduce the existing bandwidth reservation methods such as Resource Reservation Protocol (RSVP), IEEE 802.1p, and MSRP. In Section III, we propose processes for various situations related to bandwidth reservation and reservation withdrawal with using IGMP, and in Section IV, we introduce the simulation and numerical results in. Section V presents the concluding remarks.

II. RELATED WORK

A. RSVP

The Resource ReSerVation Protocol (RSVP), a kind of Integrated Service (IntServ), is a reservation mechanism executing on the transport layer. RSVP can be used by either hosts or routers to request or deliver specific levels of quality

 ISBN: 978-1-61208-203-5

of service (QoS) for application data streams or flows. RSVP defines how applications place reservations and how they can relinquish the reserved resources once the need for them has ended. RSVP operation will generally result in resources being reserved in each node along a path. RSVP does not transport application data but is rather an Internet control protocol, like ICMP, IGMP, or the routing protocol [4]. RSVP also provides receiver-initiated setup of resource reservations for multicast or unicast data flows with scaling and robustness. Figure 1 shows the reservation process via RSVP.

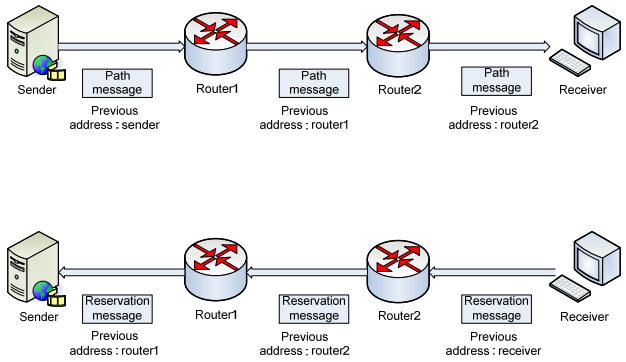


Figure 1. Reservation mechnism with RSVP

Although RSVP can control the QoS level for application data streams or flows per specific users, every device must support RSVP and store the reservation state in each node along the path [4]. The implementation of RSVP is difficult because the path of traffic is not constant on the Internet. For this reason, RSVP was little used. To solve the problem of RSVP, the Differentiated Service (DiffServ) was proposed. DiffServ provides adequate QoS via prioritizing traffic.

B. IEEE 802.1p

One type of DiffServ, IEEE 802.1p that is operated at the data link layer, processes frames according to their priority on the Ethernet [5]. IEEE 802.1p defines eight different classes of available service, usually expressed through the 3-bit priority field. The most important is priority 7 which corresponds to the network control frame, and priority 0 which corresponds to the best effort traffic that is the least important frame. If some frame belongs to the multimedia service, then its priority is 4 or 5. In this case, the multimedia frame has higher priority than other best effort frames, so multimedia frame could be guaranteed QoS. Figure 2 shows the structure of Ethernet frame using IEEE 802.1p.

However, IEEE 802.1p has two problems [6]. First, when two or more frames having the same priority arrive at the switch at the same time, some of these frames are discarded due to the limited queue size on the switch. If the time-sensitive frame is discarded, then large jitters occur and the media cannot be guaranteed QoS. Second, the more the hop count increases, the more the potential delay of time-sensitive frame increases due to the time-sensitive frame of other applications. The reason that transmission delay increases is due to accumulating delay by competition with the same priority frames and the effects of lower priority frame for non-preemption. Therefore, DiffServ is inadequate for specific applications.

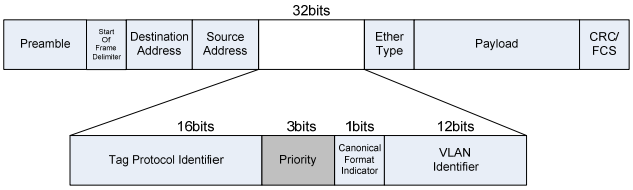


Figure 2. The frame structure using IEEE 802.1p

C. MSRP

MSRP that is being studied in IEEE 802.1 AVB TG is used in Ethernet and works by making a sub spanning tree for specific traffic [7]. MSPR reserves bandwidth on the Ethernet, so the reservation path is fixed. In addition, MSRP does not cause delays by other traffic because MSRP is a kind of IntServ.

In MSRP, a Talker means that the node can serve multimedia streaming to other nodes on the LAN, while a Listener receives multimedia streaming service from a Talker. MSRP starts a reservation when Talker announces that they can serve a multimedia service, or Listener announces that they want to receive a multimedia streaming service in the LAN. MSRP uses five types of messages. Each type is presented in Table 1.

TABLE I. MSRP MESSAGE TYPES

Message		Description
Talker	Advertise	Advertise for a stream that has not encountered any bandwidth or other network constraints along the network path from the Talker.
	Failed	Advertisement for a Stream that is not available to the Listener due to bandwidth constraints or other limitations somewhere along the path from the Talker.
Listener	Asking Failed	One or more Listeners are requesting attachment to the Stream. None of those Listeners are able to receive the Stream because of network bandwidth or resource allocation problems.
	Ready	One or more Listeners are requesting attachment to the Stream. There is sufficient bandwidth and resources available along the path(s) back to the Talker for all Listeners to receive the Stream
	Ready Failed	Two or more Listeners are requesting attachment to the Stream. At least one of those Listeners has sufficient bandwidth and resources along the path to receive the Stream, but one or more other Listeners are unable to receive the stream due to network bandwidth or resource allocation problems.

The Talker creates a Talker Advertise declaration message to announce to other nodes on the LAN and update

its MSRP table. A Talker Advertise message includes the MAC address of the Talker, declaration type, required bandwidth, etc. The bridge port 0, which receives the Talker Advertise message, then registers it in its MSRP table and sends it to other ports on the bridge. The bridge port 1 that receives the Talker Advertise message by bridge port 0 registers the message in its MSRP table and compares the requirement bandwidth of the message with its available bandwidth. If the available bandwidth is larger than the required bandwidth, then the message is forwarded to the other node. Otherwise, the message is changed to Talker Failed message and forwarded.

When A Talker Advertise message arrives at a Listener, if the Listener wants to receive the Talker's service, then the Listener generates a Listener Ready message and sends it to the Talker. A Listener Ready message has only the StreamID of the Talker; StreamID consists of the MAC address of the Talker and an integer number. Bridgeport 1, which receives the Listener Ready message, reserves the required bandwidth, if port 1 has sufficient available bandwidth. When the bandwidth reservation is successful at port 1, the Listener Ready message is forwarded to the Talker by MSRP attribute propagation. The Listener Ready message arrives at the Talker, and the Talker then compares its MSRP table with the streamID in the Ready message. If the Listener Ready message is associated with a stream that the Talker can supply, then the Talker can immediately start the transmission for this stream. Figure 3 illustrates the MSRP reservation success process.

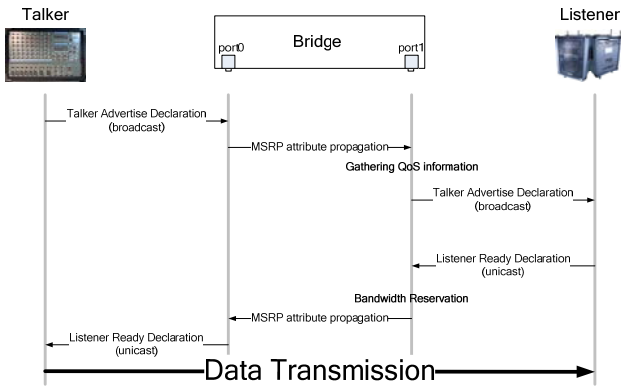


Figure 3. Bandwidth reservation success process

In the cases of reserved bandwidth withdrawal, the Talker stops streaming or the Listener declines to receive the stream of Talker service. If the Talker wants to stop the services, then the Talker generates a Talker Failed message and sends it to the Listener, who in turn stops transmitting the stream. The bridge port 0, which receives the Talker Failed message, sends to some ports that have already reserved bandwidth for the stream of the Talker. The bridge port 1 that receives the Talker Failed message by port 0 returns the reserved bandwidth and forwards the Talker Failed message. The Listener that receives Talker Failed message deletes the streamID of Talker Failed message in its MSRP table. In the case of bandwidth reservation failure, the outgoing port of bridge has insufficient bandwidth than the required bandwidth when the Talker Advertise message is forwarded. In this case, the Talker Advertise message is changed to Talker Failed message and forwarded. Then, the MSRP defines a bandwidth reservation/withdrawal and bandwidth reservation failed in the Ethernet.

However, the bandwidth reservation fail situation was able to occur in MSRP. That can happen such as the state of the listener and available bandwidth is insufficient when not only the Talker Advertise message arrives but also when the Listener Ready message arrives at the port. MSRP, which has not been researched until now, can process only a few situations in which bandwidth reservation failed, so it is not effective in a QoS guaranteed environment. When a number of services share limited resources, bandwidth reservation failure will occur; this problem must be resolved.

III. IGMP-BASED BANDWIDTH RESERVATION

The Internet Group Management Protocol (IGMP) is a communications protocol used by hosts and adjacent routers on IP networks for the purpose of establishing multicast group memberships. IGMP is used to join and leave multicast group memberships. The latest IGMP is version 3, but we use IGMP version 2, which is widely used. We suggest a bandwidth reservation using IGMP, which is a similar MSRP mechanism, and implemented a bandwidth reservation/withdrawal process on IPTV multicast group join and leave.

- In the case of terminal node send Membership Report message

When a Membership Report message sent by a terminal node sends arrives at multicast-supported router, the router starts a streaming service to the node. As a result, intermediate device receives the Membership Report message, and the device needs to reserve bandwidth. When intermediate devices receive a Membership Report message, the device checks available bandwidth at incoming port. If the available bandwidth is can support the demanded multicast stream bandwidth of terminal node, then the device reserves the demanded bandwidth and forwards a Membership Report message. When available bandwidth of the device port is insufficient, the device discards the Membership Report message. In traditional IGMP, the intermediate device receives a Membership Report message, and the device broadcasts (or multicast) the message to restrict unnecessary traffic. However, we use a Membership Report message for bandwidth reservation, and the device does not broadcast but only sends the message to the multicast router side. Therefore, another terminal node sends a Membership Report message that is about the same multicast group for bandwidth reservation. We show this process in the Figure 4.

 ISBN: 978-1-61208-203-5

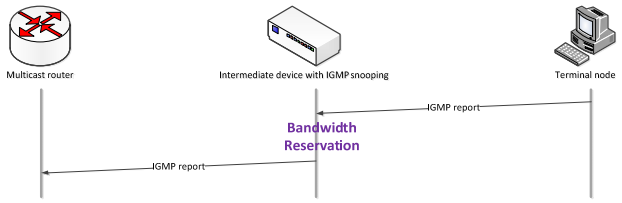


Figure 4. Bandwidth reservation process with IGMP Membership Report message

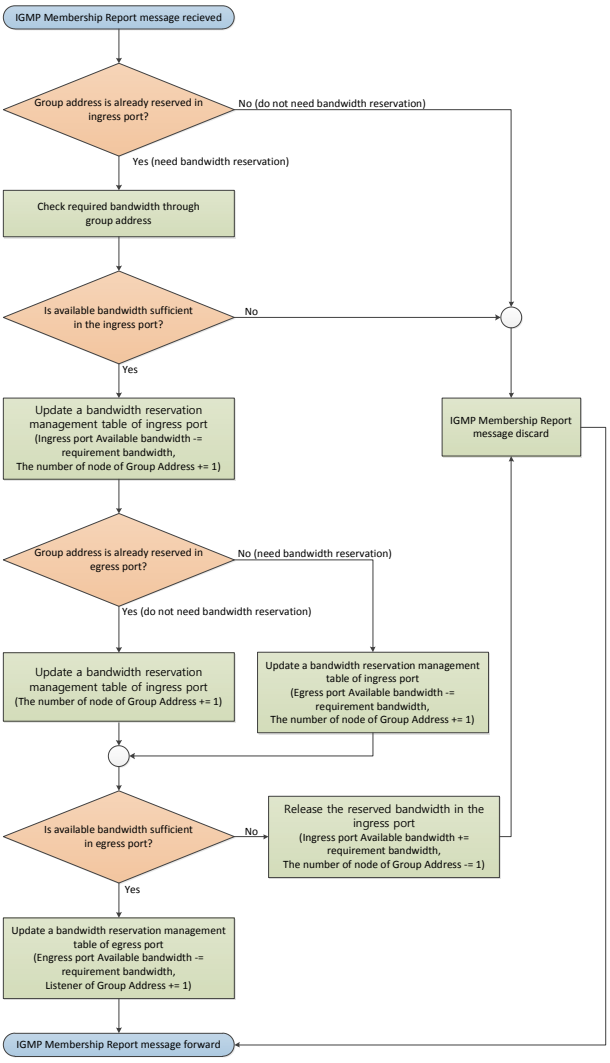


Figure 5. Bandwidth reservation flow chart in intermediate device

A bandwidth reservation process in the intermediate device is as follows. First, it checks that the group address of Membership Report message is reserved on the ingress port. When the group address is already reserved, we only update the number of terminal node of the group address. When the group address is not reserved, the device check required bandwidth of the group address, then check whether available bandwidth is sufficient or insufficient for support required bandwidth. If the available bandwidth is insufficient, then it is regarded as bandwidth reservation failure, and the device discards the Membership Report message and sends a message that notifies that the required service cannot be provided due to insufficient bandwidth to the terminal node. On the other hand, if the available bandwidth is sufficient, then the device updates the bandwidth reservation management table of ingress port and allocates the bandwidth. The device checks the bandwidth reservation management table of egress port. If the group address is not reserved in the egress port, then we sue the same process for egress port as that used for ingress port. If the group address is already reserved in the egress port, then the device forwards the message to the multicast router side. However, if the available bandwidth is insufficient in the egress port, then release the reserved bandwidth in the ingress port and discard the message. We show this process in Figure 5.

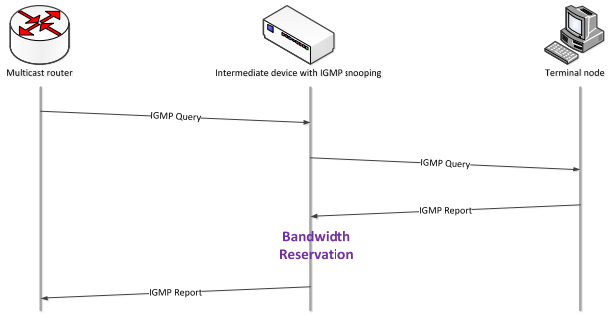


Figure 6. Bandwidth reservation process via IGMP Membership Query message

- In the case of multicast router sending the Membership Query message

In this case, a terminal node that wants to join a multicast group or already has joined a specific multicast group, sends a Membership Report message to the multicast router for bandwidth reservation. This case is the same as that of the previous case in that the terminal node sends a Membership Report message. Figure 6 is shows the bandwidth reservation process via IGMP Membership Query message

- In the case of terminal node sending a Leave Group message

In this case, the terminal node does not receive a multicast stream any more, and an intermediate device needs to release the bandwidth of multicast stream of the terminal node. We define this process as “bandwidth withdrawal”. The device forwards the Leave Group message to the multicast router side after bandwidth withdrawal. The router that receives the Leave Group message sends a Membership Query message with the group address of the Leave Group message to other terminal nodes. After this, nodes that have received the Membership Query message via multicast send a Membership Report message, and this is similar to that in the case of terminal node sending a Membership Report message. Figure 7 shows the bandwidth reservation process

 ISBN: 978-1-61208-203-5

via IGMP Leave Group message, and Figure 8 shows the bandwidth withdrawal process in an intermediate device.

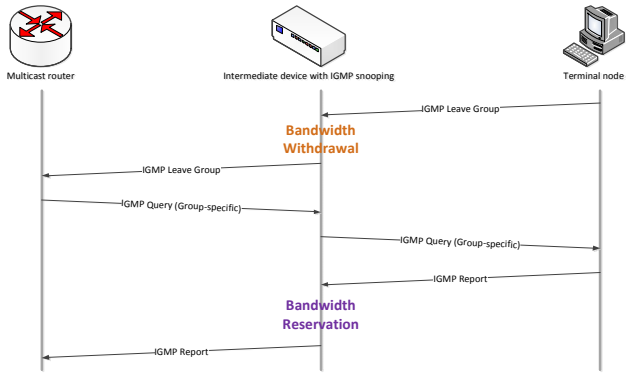


Figure 7. Bandwidth reservation process via IGMP Leave Group message

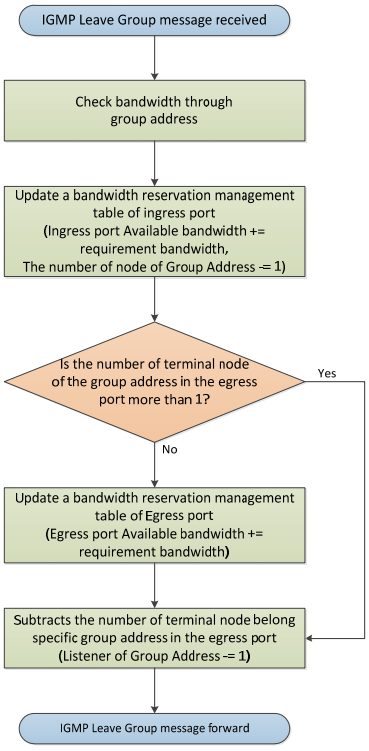


Figure 8. Bandwidth withdrawl flow chart in intermediate device

The bandwidth withdrawal process is as follows. When the ingress port of the device receives a Leave Group message, the device check the allocated bandwidth for group address in the message and then updates the bandwidth reservation management table of the ingress port and releases the bandwidth. If the number of terminal node of the group address specified in the Leave Group message is one, then the bandwidth reservation management table of the egress port also updates as according to the ingress port. If the number of terminal node of the group address in the egress port is more than one, then the device only subtracts the number of terminal node in the egress port, and forwards the Leave Group message.

IV. SIMULATION RESULT

To show the efficiency of the proposed bandwidth reservation via IGMP, we used the network simulator OPNET. We compared the traditional IGMP with our suggestion in terms of multicast streaming. We set up a network environment using 6 L3 switches that are intermediate devices, 10 terminal nodes, and the load time-sensitive traffic, non-time-sensitive traffic at each link. Further, we used time-sensitive traffic with voice and video traffic. Therefore, each node sends the Membership Report message to the multicast router at different time intervals. Figure 9 shows our network topology.

The traffic of each source is 100Mbps, and we set the outgoing stream frame size of each source to 125 Kbytes. For the analysis of end-to-end delay is according to network load. Therefore, we set different join times for each terminal node, as is shown in Table 2.

TABLE II. JOIN TIME OF EACH TERMINAL

Terminal node	Join time (sec)	Terminal node	Join time (sec)
User1	75	User5	85
User2	80	User6	90
User3	80	User7	90
User4	85		

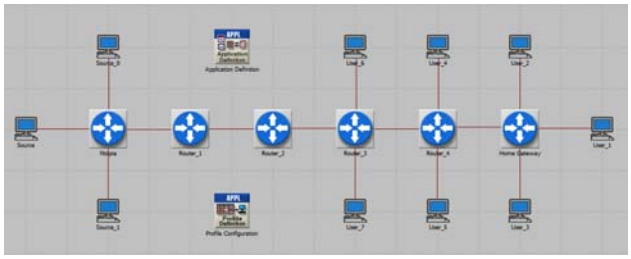


Figure 9. The network simulation topology

The traffic generation rate and characteristic follow that voice is Expedited Forwarding (EF), video is Assured Forwarding (AF), and non-time-sensitive traffic is Best Effort (BE). The EF traffic is used to transmit voice data, and it is a Constant Bit Rate (CBR) in the ATM network. The AF traffic is video data type such as MPEG-1, MPEG-2 and H.264. It is used by Video on Demand (VoD), video conferencing, etc. The AF traffic regards bandwidth with great importance, but it is less sensitive to delays than EF traffic. The BE traffic is used by legacy Internet services such as web services, e-mail and FTP. The BE traffic does not need real time transmission. Therefore, Walter et al. discovered that AF and BE traffics have self-similarity characteristic [8]. Figure 10 and Figure 11 show an end-to-end delay with traditional IGMP, and after bandwidth reservation using IGMP, respectively.

 ISBN: 978-1-61208-203-5

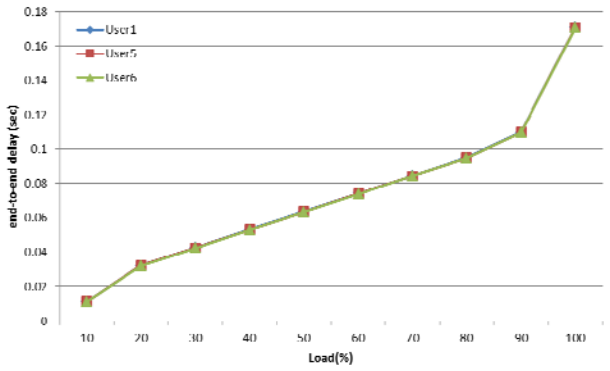


Figure 10. End-to-end delay with traditional IGMP

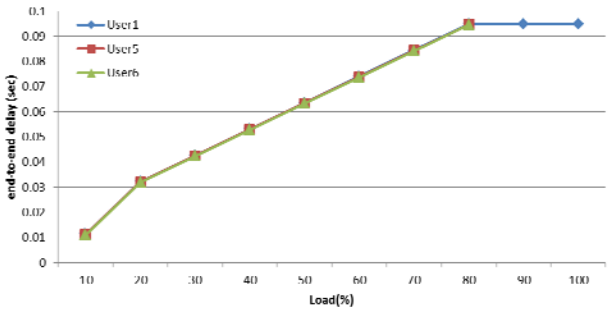


Figure 11. End-to-end delay after bandwidth reservation using IGMP

In Figure 11, the traffic node is over 90%, and multicast traffics, except User1, do not transmit because *User1* had the first IPTV streaming service after bandwidth reservation, while *User5*, *and User6* cannot IPTV the streaming service because the available bandwidth is insufficient due to bandwidth reservation by *User2 and User3*. Therefore, the QoS of User1 is guaranteed.

V. CONCLUSION AND FUTURE WORK

Nowadays, as network infrastructure and device continue to advance, various network-based applications appear among those applications, IPTV is increasing rapidly worldwide. To support IPTV services, sufficient network resources are required. This paper proposed the MSRP mechanism to reapply to the IPTV environment. We applied a bandwidth reservation mechanism using IGMP message. The bandwidth reservation is similar to that of MSRP, and we improved the network resource efficiency to process the wasted bandwidth due to bandwidth reservation failure. We simulated our proposed method via OPNET, and the network simulator confirmed that the end-to-end delay of multicast traffic is shorter than that of the traditional method.

In the following research, we expect longer channel zapping time than that of the traditional method because of the additional bandwidth process. Therefore, we need to improve channel zapping time to match that of traditional IGMP, and support suitable IPTV service. Furthermore, when wired/wireless devices join the same streaming service, a study is needed to achieve a bandwidth reservation mechanism to guarantee QoS among them.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2011-0004184), and Kyungpook National University Research Fund, 2012.

REFERENCES

[1] Yang Xiao, Xiaojiang Du, Jingyuan Zhang, Fei Hu, and Sghaier Guizani, “Internet Protocol Television (IPTV):The Killer Application for the Next-Generation Internet,” *IEEE Communications Magazine*, vol. 45, Issue 11, pp. 126-134, November 2007.

[2] Ken Kerpez, Dave Waring, George Lapiotis, J. Bryan Lyles, and Ravi Vaidyanathan, "IPTV Service Assurance," *IEEE Communications Magazine*, vol. 44, Issue 9, pp. 166-172, September. 2006.

[3] IEEE 802.1 Audio/Video Bridging Task Group Home Page, http://www.ieee802.org/1/pages/avbridges.html, Septermber 2008.

[4] Lixia Zhang, Steve Deering, Deborah Estrin, Scott Shenker and Daniel Zappala, “RSVP: A New Resource ReSerVation Protocol,” *IEEE Network*, vol. 7, pp. 8-18, September 1993.

[5] IEEE Draft Standard for Local and Metropolitan Area Networks virtual bridged local area networks – Vertual Bridged Local Area Networks, IEEE std 802.1Q, March 2006.

[6] King-Shan Lui, Whay Chiou Lee and Klara Nahrstedt, “Link Layer Multi-Priority Frame Forwarding,” *IEEE International Conference on communications*, vol. 3, pp. 1573-1577, May 2003.

[7] IEEE Draft Standard for Local and Metropolitan Area Networks virtual bridged local area networks – Amendment9 : Stream Reservation Protocol (SRP), IEEE 802.1Qat, November 2009.

[8] Walter Willinger, Murad S. Taqqu, Robert. Sherman and Daniel V. Wilson, “Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level,” ACM SIGCOMM ’95, pp. 100-113, April 1995.

 ISBN: 978-1-61208-203-5

Novel 3-Stage Scheduler for Real Time Traffic in an OFDMA System with Delay and Retransmission Constraints

Suman Kumar, Krishnamurthy Giridhar
Dept. of Electrical Engineering
Indian Institute of Technology Madras
Chennai, 600036, India
Email: {suman, giri}@tenet.res.in

Sheetal Kalyani
Centre of Excellence in Wireless Technology
IITM Research Park
Chennai, 600113, India
Email: sheetal@cewit.org.in

***Abstract*—Emerging wireless communication systems use MIMO-OFDM with adaptive modulation and Hybrid Automatic Repeat Request (HARQ) techniques to enables high bit rate and low packet error rate. In this work, a 3-Stage packet scheduling algorithm that is HARQ aware is proposed, which supports real time service with multi-level delay constraints and retransmission constraints in OFDM systems. For performance analysis, three Quality of Service (QoS) parameters namely packet loss rate, fairness, and throughput are studied. Corresponding to these three metrics and depending upon the delay and retransmission constraints, a 3-Stage scheduling strategy is proposed. It is assumed that the packets are lost due to violation of the delay constraint and/or channel induced error, even after allowing the maximum number of retransmissions. Simulation result shows that this novel 3-Stage scheduler achieves a balance between the three QoS metrics, and could therefore be preferred over Modified Largest Weighted Delay First, Proportion Fair, and Max Rate schedulers, which can not simultaneously satisfy all the three QoS metrics.**

***Index Terms*—Proportion Fair, Max Rate, MLWDF, HARQ, Throughput, Fairness, PLR**

I. Introduction

Orthogonal Frequency Division Multiple Access (OFDMA) technology enables frequency agile resource allocation where a set of sub-carriers can be allocated to a user terminal based on the scheduling logic used. Hybrid Automatic Repeat Request (HARQ) is essentially a combination of Forward Error Correction (FEC) with Automatic Retransmission Request (ARQ). The Third generation partnership project Long term evolution (3GPPLTE) uses OFDM with link adaptation and HARQ techniques to enable low packet error rate. Therefore, it is essential to have a HARQ packet scheduler. In some of the early work on wireless packet schedulers, errors in wireless transmission have not been considered. However, unless HARQ attempts are given some priority by the schedular, it is well known that overall performance of the application would suffer [1], [2].

The Max Rate rule schedules those user whose channel condition (In interference limited deployment typical in reuse 1 OFDMA cellular system, it is the post processing Signal to Interference Noise Ratio that is used in deciding which user gets scheduled. For brevity, we refer this here simply as channel condition.) is better than the other users and thereby maximizes the throughput. A user whose channel condition is bad, gets scheduled rarely and thus Max Rate does not guarantee fairness. To increase the Fairness among users Proportional Fair (PF) scheduler [3] was proposed. It schedules the user by comparing the ratio of current data rate to average data rate of a particular user so that fairness can be addressed. Along with fairness and throughput, packet delay is a key parameter to measure the performance of highly delay sensitive, real time application like video streaming. PF scheduler as such does not consider packet delay. Modified Largest Weighted Delay First (MLWDF) algorithm [4] is able to handle delay sensitive traffic well. It schedules the user by comparing the combination of packet delay, current data rate and average data rate in an optimal way.

In this paper, we consider all the three Quality of Service (QoS) parameters namely: (i) packet loss rate (PLR), (ii) fairness and (iii) sum throughput. Corresponding to the three metrics we propose a scheduling strategy which has three stages. From the quality of service (QoS) perspective of real time traffic, it is essential to give priority to minimize packet loss rate (PLR), maximize fairness and maximize throughput simultaneously that is what this 3-Stage scheduler aim to do. The 3-Stage scheduler is compared for real time traffic with the Max Rate, the PF, and the MLWDF algorithm. Using simulation results we show that the proposed scheduler gives a good compromise between the three QoS metrics.

The remainder of the paper is organized as follows. In Section 2, the LTE-like simulation model are presented that is considered in this paper. Section 3 discusses about the 3-Stage scheduler followed by simulation results are shown in Section 4, while conclusion are given in Section 5.

II. Lte-Like Simulation Model

Fig. 1. depicts the downlink frame structure LTE Standard. On frequency axis total bandwidth is divided into N sub bands and in time axis into transmission time interval (TTI) each with length of 1 ms. Here 1 sub band contains 3 physical resource

 ISBN: 978-1-61208-203-5

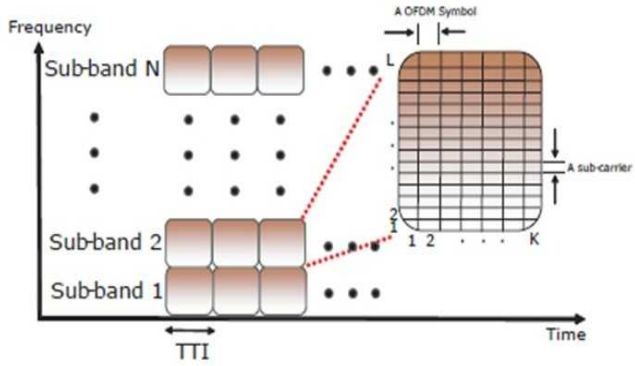


Fig. 1. Shows OFDMA framing and channelization

block (PRB) and 1 PRB has 12 sub-carriers and 14 OFDMA symbols. The simulation of 3-Stage, MLWDF, PF and Max Rate have performed using seven hexagonal cells of 10 MHz bandwidth with 50 PRBs and 2 GHz carrier frequency. There are 3 sectors in a cell and each sector contains 10 users. Users are uniformly located within the cell and constantly moving at a constant speed of 8.33 Km/h in random directions. It is assumed that each user reports its instantaneous downlink SNR values on each PRB and at the beginning of each TTI to the serving node. The reported instantaneous downlink SNR value is used to determine the feasible data rate for one PRB. The 3GPP LTE Downlink system parameters are given in Table I.

An incremental redundancy HARQ protocol is used. The HARQ process takes 2ms, i.e., 2 TTI round trip time and maximum number of retransmission is limited to 4. It is assumed that the scheduling interval is 1 TTI i.e. 1ms and the number of PRBs that may be allocated to a user in each scheduling interval is variable.

MLWDF [1], PF [4], Max Rate [4] algorithms are proposed for the single carrier transmission. We modified these algorithms to support multi-carrier transmission in the downlink LTE system. In this simulation, retransmissions takes place for each algorithm, if transmission of any packet fails. The following QoS in equations (1), (2), (3) are used for performance analysis.

$$Average\ throughput = \frac{1}{N*T}\sum_{i=1}^{N}\sum_{t=1}^{T} t\_put_i(t) \quad (1)$$

$$PLR = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} pd_i}{\sum_{i=1}^{N}\sum_{t=1}^{T} pa_i} \quad (2)$$

$$Fairness = \frac{1}{N}\frac{(\sum_{i=1}^{N} t\_put_i)^2}{\sum_{i=1}^{N}(t\_put_i)^2} \quad (3)$$

where $t\_put_i(t)$, pd_i , pa_i are the throughput, total size of discarded packets and the total size of all packets that have

TABLE I
SIMULATION PARAMETERS

Parameters	values
Carrier Frequency	2 GHz
Bandwidth	10 MHz
Number of Sub-carriers	600
Number of PRBs	50
Number of Sub-carriers per PRB	12
Slot Duration	1ms
Scheduling Time(TTI)	1ms
Number of OFDMA Symbols per Slot	14
FFT size	1024
HARQ scheme	Incremental redundency
Maximum Allowed retransmission number	4
Total number of User	210
Number of Interferer cell	2

arrived into the buffer of user i at time t respectively. T and N are the total number of slots for which simulation has done and total number of users respectively. Referred to [5].

III. DESCRIPTION OF 3-STAGE SCHEDULER

3-Stage scheduler divides the users into three stage based on how close the packets are from transmission deadline.
TFT_i= Time for Transmission [2] is defined as the time duration up to which packet i can stay in the buffer for transmission. It has an integer value normalized by TTI duration.
$TFL_i(n)$ = Time duration upto which packet is not dropped. Equations (4) and (5) shows analytical description of TFT_i and $TFL_i(n)$.

$$TFT_i = k * TTI \quad (4)$$

$$TFL_i(n) = TFT_i - W_i(n) \quad (5)$$

Where $W_i(n)$ is waiting time for a Head of Line (HOL) packet in ith buffer in nth TTI and k is an integer. The proposed 3-Stage scheduler has following steps:

Step A. Divide the users into three stages depending upon the value of K_{max} (maximum value of k among all user) and $TFL_i(n)$ such that distribution of $TFL_i(n)$ along the stages is in Geometrical Progression (GP). Stage 1 contains the users whose $TFL_i(n)$ is one. Stage 2 contain the users whose $TFL_i(n)$ value are 2, 3 or 2, 3, 4 or 2, 3, 4, 5 (depending upon the value of k_{max}) and the remaining users will be in stage 3. Distribution of users for different value of K_{max} is given in Table II, users will be distributed similarly for higher values of K_{max}. GP is used here, because it gives a good compromise

 ISBN: 978-1-61208-203-5

TABLE II
DISTRIBUTION OF USERS BASED ON TFL

	TFL	TFL	TFL
stage1	1	1	1
stage2	2, 3	2, 3, 4	2, 3, 4, 5
stage3	4, 5, 6, 7	5, 6....13	6, 7...21
	$K_{max} \leq 7$	$8 \leq K_{max} \leq 13$	$14 \leq K_{max} \leq 21$

TABLE III
EXAMPLE OF A SUB-BAND ALLOCATING IN STAGE 1

	user 1	user 2	user 3
subband 1	5	6	7
subband 2	7	6	5
subband 3	7	5	6
Allocated band	Subband 3	Subband 2	Subband 1

among the three QoS metrics as shown by extensive simulation results, some of which are shown in section IV . Instead of GP, other progressions can be used to divide the users into stages depending on the QoS requirement. For example if we keep more number of users in the third stage compared to what GP provides, it will give more throughput than GP gives. However this increase will come at the cost of higher packet loss rate and degraded fairness. Similarly if we put more number of users into second stage performance of fairness may improve at the cost of degradation of the performance of packet loss rate and throughput.
In summary the purpose of the three stages are: -
Stage 1: To minimize the Packet Loss Rate.
Stage 2: To maximize the Fairness.
Stage 3: To maximize the Throughput.

From the QoS perspective of real time traffic it is essential to give priority to minimize the packet loss rate, maximize the Fairness and maximize the throughput respectively therefore we schedule stages 1, 2 and 3 respectively.

Step B. Schedule the users of stage 1:- Consider the channel matrix, where each column corresponds to the different user of stage 1 and each row corresponds to a sub band. C_{ij} denotes the number of bits can be transmitted through sub band i to the user j. Depending upon the modulation and coding scheme levels, C_{ij} will have N different Number of bits values i.e. NOB_1, NOB_2, NOB_3, NOB_4.....NOB_N. Define coordinate of each C_{ij} which is defined as (s_i, u_j) and coordinate group of NOB_i i.e CG_{NOB_i} which contains coordinate of all C_{ij} which is equal to NOB_i. Assuming $NOB_1 < NOB_2 < NOB_3 < NOB_4 <.....< NOB_N$, then to improve QoS CG_{NOB_N} must be allocated prior to the $CG_{NOB_{N-1}}$ and so on. Among the C_{ij} in coordinate group CG_{NOB_i}, C_{ij} is allocated when its coordinate is unique in that particular group. Otherwise choose an unique combination from all possible combinations such that sum throughput is maximized and allocate them.
Here in Table III NOB_3 is 7, NOB_2 is 6 and so on. Coordinate Groups are defined by
CG_{NOB_3}= {(s_1,u_3),(s_2,u_1),(s_3,u_1)}
CG_{NOB_2}= {(s_1,u_2),(s_2,u_2),(s_3,u_3)}
To minimize the PLR CG_{NOB_3} should be allocated first. Since only coordinate(s_1,u_3) is unique so subband 1 is allocated to user 3. Now user 1 can be served by subband 2 or subband 3 so following are the possible combinations
combination 1: (s_2,u_1) + (s_3,u_2) = 12
combination 2: (s_2,u_2) + (s_3,u_1) = 13
choose combination 2 since sum throughput is maximum and allocate them.

Step C. Schedule the users of stage 2:-
Schedule the users of stage 2 when its C_{ij} is more then C_{avg} in round robin fashion so that fairness can be maximized.

$$C_{avg} = \frac{1}{M*N} \sum_{i=1}^{M} \sum_{j=1}^{N} C_{ij}$$

Where N = no of users in stage 2 and
M = number of sub bands left.

Step D. Schedule the users of stage 3:
Among all the users of stage 3 schedule the user whose C_{ij} value is equal to NOB_N. There may be a case where a user contains more than one C_{ij}, whose value is equal to NOB_N in that case we schedule the user more than one time provided available bits for transmission are sufficient. Again we schedule the user whose C_{ij} value is equal to NOB_{N-1} followed by NOB_N and so on. This process will continue till sub-band left.

IV. SIMULATION RESULT AND DISCUSSION

We simulate downlink physical layer Long Term Evolution release 8 [6] along with the parameter mentioned in Section II. Four types of real-time traffic sources with different delay constraints are used, as given in Table IV. Fig. 2. shows PLR performance of 3-Stage, PF, MLWDF and Max Rate with the increasing number of arrival bits. 3-Stage scheduler improves PLR performance by giving more priorities to the HARQ users and the users closing to transmission deadline. A significant degradation of the PLR performance in the Max Rate and PF are because they do not consider delay of packets. As MLWDF consider waiting time of packet so PLR of MLWDF is just followed by 3-Stage scheduler.

Fig. 3. shows average user throughput of the 3-Stage, PF, MLWDF, and Max Rate. From the figure, it can be observed that among the PF, MLWDF, Max Rate, and 3-Stage, Max Rate achieves the highest throughput as it consider only channel

 ISBN: 978-1-61208-203-5

condition. The proposed scheduler gives throughput slightly lower than Max Rate but higher than PF. MLWDF returns the lowest throughput as it consider both channel condition and waiting time of a packet however all four schedulers achieve almost same throughput performance at a higher number of arrival bits per frame.

The fairness performance of each scheduling algorithm is shown in Fig. 4. It can be observed that the 3-Stage scheduler achieves a fairness almost equal to that of PF and MLWDF. The Max Rate returns the lowest fairness as it only considers the best channel condition for the scheduling decision.

TABLE IV
TFT FOR FOUR REAL-TIME TRAFFIC SOURCES

Service	Traffic class	TFT
Voice	Conversational	200ms
Gaming	Conversational	400ms
Audio streaming	Streaming	1500ms
Video streaming	Streaming	2000ms

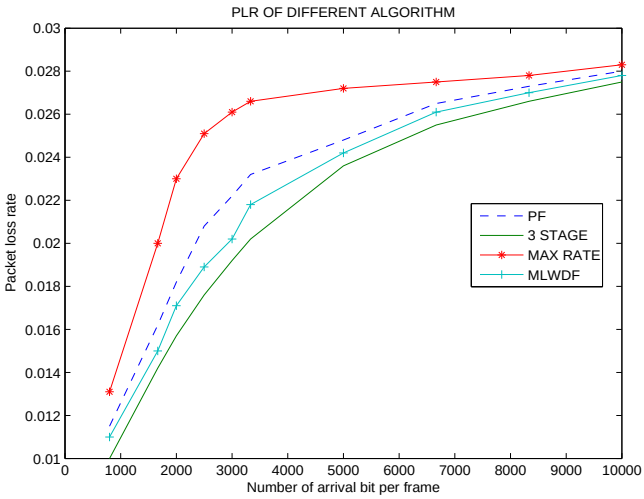


Fig. 2. Shows plr vs. no of arrival bit per frame.

V. CONCLUSION

This paper investigated the performance of proposed algorithm along with well known scheduling algorithm in the downlink 3GPP LTE system. Using the TFL parameter, an efficient scheduling scheme that always prioritizes urgent real time traffic users and HARQ users in OFDMA environment has developed. Simulation results have shown that the proposed scheduler scheme gives a good compromise among the three QoS metrics.

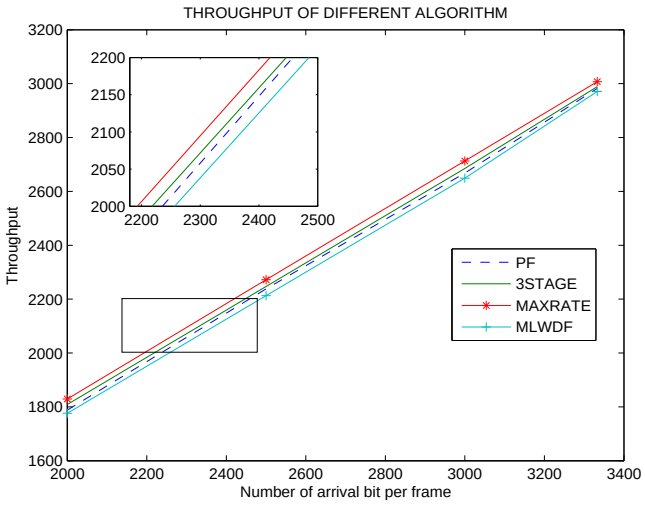


Fig. 3. Shows Average throughput vs. no of arrival bit per frame.

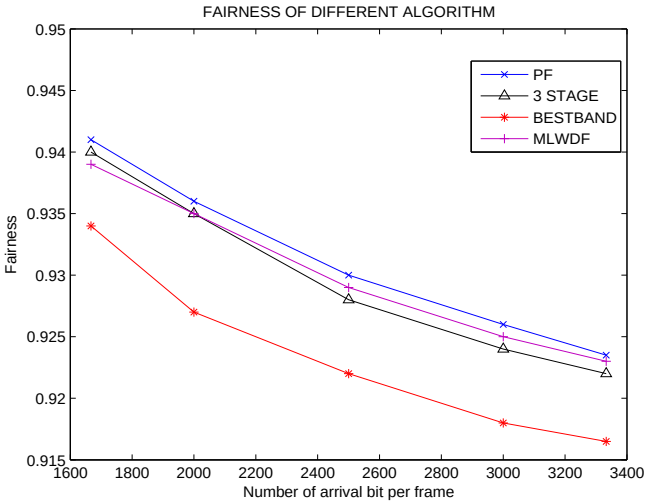


Fig. 4. Shows fairness vs. no of arrival bit per frame.

REFERENCES

[1] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, vol. 39, no. 2, pp. 150–154, feb 2001.

[2] J. Park, S. Hwang, and H. Cho, "A packet scheduling scheme to support real-time traffic in OFDMA systems," in *IEEE 65th Vehicular Technology Conference, 2007. VTC2007-Spring.* IEEE, 2007, pp. 2766–2770.

[3] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *IEEE 51st Vehicular Technology Conference Proceedings, 2000. VTC 2000-Spring Tokyo*, vol. 3. IEEE, 2000, pp. 1854–1858.

[4] H. Ramli, R. Basukala, K. Sandrasegaran, and R. Patachaianand, "Performance of well known packet scheduling algorithms in the downlink 3GPP LTE system," in *2009 IEEE 9th Malaysia International Conference on Communications (MICC).* IEEE, 2009, pp. 815–820.

[5] H. Ramli, K. Sandrasegaran, R. Basukala, and T. Afrin, "HARQ aware scheduling algorithm for the downlink LTE system," in *2011 4th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO).* IEEE, 2011, pp. 1–4.

[6] Long Term Evolution Release 8, "www.3gpp.org/LTE."

 ISBN: 978-1-61208-203-5

Application Aware Mechanisms in HSPA Systems

Péter Szilágyi and Csaba Vulkán
Nokia Siemens Networks
Budapest, Hungary
{*peter.1.szilagyi, csaba.vulkan*}*@nsn.com*

***Abstract*—Nowadays, the focus of network operation is no longer on the coverage and basic services but it is rather centered around the quality of experience that can be provided to the subscribers and on the capability of smoothly operating complex, interactive and increasingly data hungry applications on mobile platforms. Despite the increased system capacity, improved efficiency and sophisticated quality of service (QoS) architectures, the right level of quality of experience (QoE) requires application level differentiation. Using a single data bearer for each user equipment (UE), which is a common setup due to system and equipment limitation and the bearer centric QoS architectures, represents a barrier in providing true differentiation between simultaneously used applications. This paper discusses possible application level differentiation mechanisms either assuming a single data bearer per UE or utilizing the potential of secondary bearers to prioritize selected applications. The mechanisms were evaluated and compared by simulations focusing promoting web browsing over bulk data transfer in a High Speed Packet Access (HSPA) network. Results show that application differentiation mechanisms are able to significantly improve the quality of experience.**

***Keywords*-application awareness; HSPA; quality of experience; quality of service; simulation and modeling**

I. INTRODUCTION

The increasing prevalence of mobile devices with enhanced capabilities of running multimedia and web-based applications requires network-side evolution to fully serve the traffic demand. The nowadays spreading smart phones give access to the full spectrum of Internet-based applications already familiar from desktop computers, such as streaming multimedia, web browsing, mail, instant messaging, micro blogging, etc., encouraging the usage of multiple applications and services at the same time. A natural expectation of the users is to have reasonably good access to all applications even if they are run simultaneously, regardless of their different QoS requirements. However, despite the increased system capacity, high data rates and low latency provided by the evolved systems such as HSPA [1] and Long Term Evolution (LTE), there are still not enough resources in the mobile networks (especially considering the capacity limited last mile) to be able to smoothly support this user behavior without active QoS management on the network side. The end-user quality of experience greatly depends on how well the network is able to fulfill the QoS requirements of the applications [2]. Currently, due to network and equipment limitations, the entire data traffic of the users generated by the various applications is served by one data bearer. The QoS architecture is bearer centric, therefore all applications of the user receive the same service regardless of their different quality requirements; this makes it difficult for operators to enforce policies such as separately demoting bulk traffic or promoting premium services or applications. A possible solution is to use application aware mechanisms that are able to provide differentiation among the simultaneous applications run by the users. The requirement for application aware QoS has been raised not only in mobile networks [3] but in the context of transport network provisioning as well [4]. Research towards enhancing QoE is important not only in future network architectures such as LTE and beyond but also in HSPA networks, which today serve the vast majority of mobile broadband users.

In HSPA systems, bearers are used to deliver traffic according to a predefined set of QoS parameters over the radio access network (RAN) between the UE and the Radio Network Controller (RNC), referred to as the radio access bearer (RAB) service, and between the RNC and the core network (CN), referred to as the CN bearer service. A one-to-one mapping between RABs and CN bearers is done at the RNC to provide the Universal Mobile Telecommunications System (UMTS) bearer service [5]. At bearer setup, the UE can request certain QoS parameters such as guaranteed bit rate (GBR) or traffic class (TC). Based on that and operator policy settings, the Gateway GPRS Support Node (GGSN) determines additional parameters such as traffic handling priority (THP) and allocation/retention priority (ARP) and signals them to the RNC. The RAB specific QoS parameters, such as scheduling priority indicator (SPI) and discard timer (DT) are set by the RNC based on a mapping provided by the network operator or equipment vendor and signaled to the Node B along with the GBR [6]. The GBR parameter defines the target average bit rate that the air interface packet scheduler at the Node B should try to guarantee to the bearer. The SPI (an integer taking values from the range 0–15) specifies the priority of the data flow served by the bearer. DT gives the maximum allowed waiting time of the flow's packets (before being discarded) at the Node B buffers. These parameters are used by the Node B packet scheduler upon scheduling decisions. Once the active bearers receive their GBR, the packet scheduler is supposed to distribute the remaining air interface resources by considering the priority

 ISBN: 978-1-61208-203-5

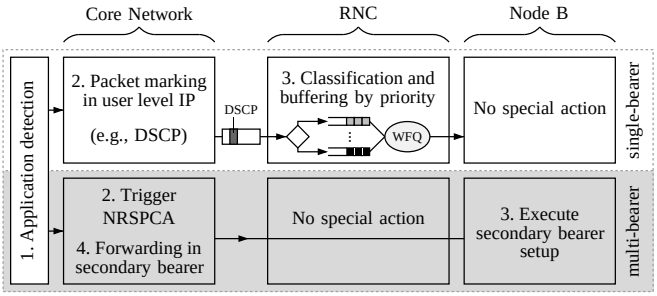


Figure 1. Concept of application aware mechanisms showing the entities involved in the discussed single- and multi-bearer alternatives.

of the bearers. An efficient packet scheduling discipline is the one implemented by the Proportional Fair with Required Activity Detection (PF-RAD) [7] scheme, that is capable of scheduling the bearers based on their weight. Accordingly, in addition to the GBR, a pre-configured parameter, the weight of the SPI class (wSPI) can be used, which is configured in the Node B for each SPI (not signaled from the RNC). Throughout this paper, we assume that the packet scheduling is based on the PF-RAD scheme. This QoS mechanism is not application aware as it is only able to differentiate among RABs but not among applications. In order to improve the situation, a couple of network-side techniques can be used, including single-bearer and multi-bearer mechanisms. Single-bearer means that the one data bearer per UE limitation is kept but the bandwidth available to the RAB is split between the applications in a predefined ratio (referred to as *in-bearer prioritization*), as proposed in [3]. Multi-bearer means to map the packets of applications with different QoS requirements to separate radio access and CN bearers, facilitated by the secondary PDP context activation procedure [8] standardized by the 3rd Generation Partnership Project (3GPP). This of course requires support from both device and network side.

In this paper, we discuss in-bearer prioritization and network requested secondary PDP context activation (NRSPCA) in detail, study their advantages and disadvantages and evaluate them based on web browsing user experience by conducting simulations in a HSPA network. The concept of single- and multi-bearer solutions is shown in Fig. 1. Results indicate that both mechanisms can considerably help enhance web page download performance; the apparatus required to implement the features can be the key differentiator in choosing the one selected for practical adoption in a real network.

The rest of this paper is organized as follows. In Section II, in-bearer prioritization is discussed. Section III provides an overview of NRSPCA and the related apparatus. In Section IV, the simulation models used in the performance evaluation of the proposed mechanisms are presented and Section V contains the simulation results and their interpretation. Finally, Section VI concludes the paper.

II. In-bearer Prioritization

The rationale behind single-bearer mechanisms is to maintain compatibility with such UEs and network-side implementations that are only capable of managing one data bearer per UE but still improve the QoE when different applications are simultaneously run by a user.

A plausible single-bearer mechanism capable of prioritizing traffic in the RAB is to mark packets in the CN according to the priority of the generating application and use per-priority packet buffering for each UE in the RNC Packet Data Convergence Protocol (PDCP) layer. The different buffers are served by Weighted Fair Queuing (WFQ) scheduler with configurable weights for the different priorities [3]. This feature requires an application detection facility in the CN, suitably in the GGSN as this is the node capable of intercepting packets arriving from external networks such as the Internet and investigate their application level content. One possible realization of application detection is Deep Packet Inspection (DPI), which can examine the TCP/IP headers of the user traffic and (or) apply pattern detection to recognize different applications. The result of the detection needs to be conveyed to the radio node where the RAB is originated, i.e., to the RNC, where the in-bearer prioritization takes place. The RNC is the best choice also as the next entity capable of accessing the application level data is the UE itself. Propagation of the application from the GGSN can be implemented by mapping the detected applications to priorities and marking the downlink (DL) packets accordingly, e.g., by utilizing the 6-bit DiffServ Code Point (DSCP) field of the inner IP header. The marking is thus encapsulated by GPRS Tunneling Protocol (GTP) and remains hidden from the transport mechanisms on the Gn and Iu-PS interfaces. For priority mapping, the following three levels may be used: middle priority for default traffic, i.e., traffic corresponding to the original QoS settings of the bearer; high priority for traffic to be prioritized; and low priority for traffic to be deprioritized. Generalization to additional priority levels is also possible.

In the RNC, the DL data packets are classified based on their priority marking and transferred to the corresponding per-priority PDCP buffer of the RAB. The amount of data sent from a given PDCP buffer to the RLC layer is determined by the WFQ mechanism and it is proportional to the weight of the buffer. The apparatus required by the mechanism is shown in Fig. 2. The solution is flexible as it allows the definition of different weight for each SPI class.

In-bearer prioritization can only be applied to the DL traffic as it is based on classification and WFQ scheduling mechanisms implemented at the RNC, where packets are multiplexed into the RAB. Such mechanisms are difficult to implement for uplink (UL) traffic as the other end of the bearer is at the UE. No other network element in between the UE and the RNC has access to the applica-

 ISBN: 978-1-61208-203-5

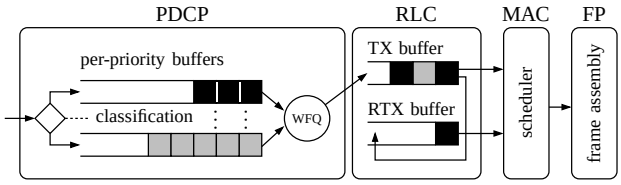


Figure 2. Implementation of in-bearer prioritization in the RNC radio protocol stack.

tion level, therefore UE side extra functionality would be required. Accordingly, network-side in-bearer prioritization transparent to the UE is feasible only in DL. The benefit of this solution is that it requires differentiation only at the radio network layer without involving any of the underlying transport network functionalities. This keeps the solution simple as the transport parameters and QoS mapping of the radio access and CN bearers are kept unchanged during their lifetime, thus no extra signaling is required.

III. Secondary PDP Context Activation

In the previous section, a single-bearer mechanism for application differentiation was discussed; in this section, a multi-bearer mechanism is presented that requires the use of multiple data bearers per UE. In the context of NRSPCA, a data bearer means both the radio access and the CN bearer, not only the RAB as with in-bearer prioritization. Multiple data bearers enable clean differentiation of applications with diverse QoS requirements without need to change the bearer-based QoS architecture. Additional bearer setup in 3GPP networks (such as HSPA or LTE) is a standardized procedure called secondary Packet Data Protocol (PDP) context activation, which can be requested either by the UE or by the network. Since application awareness requires that the network reacts to different applications of users, the secondary PDP context has to be requested by the network.

According to the 3GPP specifications, there is a PDP context for each data bearer that stores the service or Packet Data Network (PDN) the user connects to (e.g., the Internet); the IP address the UE uses in that PDN; and the QoS settings that apply to the PDP context. There are two types of PDP contexts: primary and secondary. Each different PDN to which a user is connected has an associated primary PDP context with default QoS profile attributes set according to operator policy. Users may have multiple active primary PDP contexts, one for each different PDN they connect to; however, the QoS profile of each PDP context applies to all traffic sent to or received from the corresponding PDN, i.e., although the access of different PDNs may be configured with different QoS settings, there is no means to further differentiate between traffic mapped to the same primary PDP context. The requirement of finer QoS configuration is the key motivation behind secondary PDP contexts, which allow QoS differentiation for applications and services (e.g., web browsing, FTP, P2P, streaming) over the same PDN, i.e., the Internet. Each secondary PDP context is associated with a primary PDP context, from which the PDN itself and the IP address of the UE are reused but the QoS profile can be different. A primary PDP context may have multiple secondary contexts assigned to it. Each PDP context, either primary or secondary, has a separate data bearer consisting of a RAB and a CN bearer for user plane data, i.e., the QoS configuration of the bearer is applied not only on the RAN (as with in-bearer prioritization) but consistently on the CN as well, both in UL and DL directions, which gives opportunity to prioritize applications end-to-end. Additionally, as the solution is compliant with the RAB-based QoS architecture, mapping of the bearers to the transport QoS is straightforward, which results in a compact harmonized end-to-end application aware QoS architecture.

The mapping of user-plane traffic to a certain PDP context is based on the Traffic Flow Templates (TFT). A TFT is created dynamically when a PDP context is activated and defines what kind of traffic belongs to the new context based on filters that can match, e.g., the IP address of the remote server or the source and destination TCP/UDP ports. DL TFTs are used in the GGSN for mapping DL user data to the correct GTP tunnel whereas UL TFTs are used in the UEs to implement the mapping in the opposite direction.

According to the standardized procedure of Network Requested Secondary PDP Context Activation [8], the GGSN triggers the UE to initiate the Secondary PDP Context Activation procedure with the QoS parameters and UL TFT specified by the GGSN in the first message. Thus, a functionality located in the CN is able to trigger the setup of secondary bearers with a predefined QoS configuration and, what is also important, the mapping of UL traffic can also be specified by the network. Using NRSPCA as an application aware feature is possible in a way that in case an application is detected in the CN (possibly via the same DPI mechanism also used for single-bearer mechanisms) that should be prioritized according to operator policy (e.g., HTTP traffic), the NRSPCA procedure is initiated to establish a secondary bearer with the desired QoS settings and the corresponding DL and UL TFTs are created in order to map the traffic belonging to the application into the new secondary bearer. Since the UL TFT is also created by the network and signaled to the UE, NRSPCA is suitable for treating both DL and UL traffic in a uniform way. In either case, the detection is done by the DPI located in the CN. After the application that triggered the NRSPCA finishes, which can be noticed, e.g., by activity detection, the secondary bearer should be terminated.

IV. Simulation Models for Evaluation

In-bearer prioritization and NRSPCA were evaluated by examining web page downloads in a simulated multi-cell HSPA network. The radio network layout consisted of a central cell surrounded by six other cells placed at 250 m

 ISBN: 978-1-61208-203-5

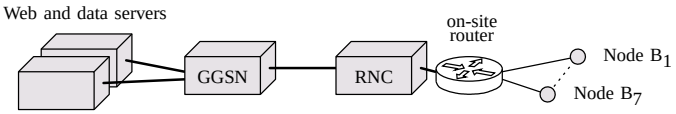


Figure 3. Simulation topology

inter-site distance. Users were distributed in the 7 HSPA cells, moving at an average of 3 kmph according to the random way-point mobility model. Wideband Code Division Multiple Access (WCDMA) air interface and handover procedures between cells were modeled in detail. The simulation topology is shown in Fig. 3; on the Iub interface, the Iub/IP/Ethernet protocol stack was used with different capacity configurations of 5, 10 and 100 Mbps, covering the range from heavy to no Iub congestion. Base stations were connected to the RNC in a star topology and implemented a Congestion Control (CC) algorithm [9].

Applications modeled in the simulation were TCP-based bulk data transfer and web browsing, the latter consisting of the complete HTTP/1.1 [10] protocol suite including Domain Name System (DNS) queries over UDP for name resolution. That is, users were executing file downloads and web page retrievals during the simulation.

The web browsing quality of experience was studied through the two most prominent quality measures: responsiveness, measured by web page download latency; and speed, measured by the page download rate. The download latency was the time between the user sending the request and receiving the first data byte of the web page. The download rate is the aggregated rate of TCP connections measured over the interval between receiving the first data byte of the main page and the download completion. Web surfing was implemented so that a random web site was selected from the list of top web sites [11] such as Google, Facebook, Wikipedia, etc., and the objects of a web page from that site were downloaded. For the simulation of web traffic, a profile was built for each modeled web page to record its main page size, the number and size of embedded objects and the server name for each object (in order to decide whether a DNS query was needed before establishing a TCP connection to the server). After a page had been downloaded, there was a random reading time in which no web traffic was generated. Then, the user visited another page from the same or another randomly chosen site.

For the sake of simplicity, two distinct user behaviors were simulated: background users having one bulk data transfer of continuous data download and multi-flow users with a similar bulk data transfer and additional web surfing. At the start of a simulation, there were 11 background users in each cell (total of 77 background users in the system) and there was one multi-flow user in the central cell. This setup was created in order to show the maximum achievable gain. With more multi-flow users, the gain would be smaller due to the increased amount of concurrent HTTP connections;

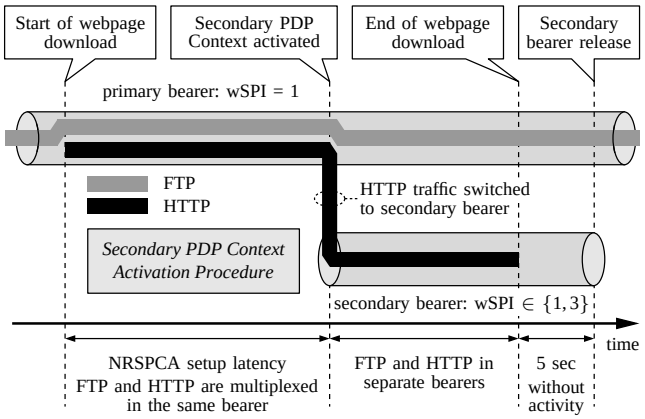


Figure 4. Modeling of Network Requested Secondary PDP Context Activation (NRSPCA) in the simulations.

however, relatively better service can be still provided to the HTTP connections since they are prioritized over the parallel applications.

For in-bearer prioritization, the DPI functionality was modeled in the GGSN so that DL packets were marked according to the application: HTTP packets were high and FTP packets were middle priority. DPI was assumed to be perfect so that all packets were marked correctly according to the corresponding application. This is possible as the DPI mechanisms available are efficiently detecting the applications with practically no latency, thus even the first HTTP packet can be treated according to the predefined QoS differentiation strategy. In-bearer prioritization was implemented in the RNC according to Section II with WFQ weights configured so that $w_{\text{high}} : w_{\text{middle}} = 9 : 1$. Due to the specified marking of HTTP and FTP, the weight of the applications was also $w_{\text{HTTP}} : w_{\text{FTP}} = 9 : 1$.

On the transport, all user-plane traffic was mapped to the same Per-Hop Behavior (PHB) group. On the radio, the default wSPI of all bearers was set to 1. There was no GBR configured to any of the bearers.

The simulation model of NRSPCA is illustrated in Fig. 4. In this case, the application detection was also done by the DPI in the GGSN but instead of marking the packets according to the application, the DPI triggered the activation of a secondary PDP context whenever a starting web page download was detected. Once the secondary bearer was established, the HTTP traffic sent in either DL or UL was mapped to that new bearer, whose wSPI was either set to 1 (i.e., the same QoS profile was used for primary and secondary bearers) or 3 (i.e., a new, better QoS profile was used for the secondary bearer in order to prioritize it over the other bearers). The signaling messages of the secondary PDP context activation were not simulated; instead, when a HTTP packet was detected for a user in the gateway, a timer was started that modeled the NRSPCA latency, i.e., the time required for completing the NRSPCA Procedure. When the timer expired, a secondary bearer was created for the HTTP

 ISBN: 978-1-61208-203-5

Table I
SHORT LABEL AND DESCRIPTION OF SIMULATION SCENARIOS

label	description
ref	reference case (no application aware feature)
wfq	in-bearer prioritization
nrspca-Z nrspca-L	NRSPCA with secondary bearer wSPI = 1; '-Z' denotes zero NRSPCA latency; '-L' denotes random NRSPCA latency between 0.8 – 1 seconds
nrspca-Z-pro nrspca-L-pro	NRSPCA with secondary bearer wSPI = 3; the meaning of '-Z' and '-L' are the same as with nrspca

packets. During the NRSPCA setup, HTTP packets were multiplexed with FTP in the primary bearer. After the web page download was complete, which was detected in the GGSN as 5 s without activity in the secondary bearer, the release of the secondary bearer was triggered.

Since the transmission of HTTP packets on the Iub interface is slower in the primary bearer during the NRSPCA setup than later in the dedicated secondary bearer (as the simultaneous applications of the users are still competing for the resources during this time), the first HTTP packets sent in the secondary bearer may arrive at the UE earlier than some of those sent through the primary bearer, potentially causing out-of-order delivery at the UE that would eventually trigger TCP Fast Recovery mechanism at the sender. In case the secondary bearer is prioritized as well, this effect is even more pronounced. In order to prevent this potential problem, we propose that after the secondary bearer setup is complete, the GGSN sends an end marker (GTP-U packet) [12] in the primary bearer and starts forwarding subsequent DL HTTP packets in the secondary bearer. Packets received in the secondary bearer are buffered at the RNC until the end marker arrives; on that occasion, the RNC transmits all packets it has buffered in the secondary bearer in the order of their arrival. After that, subsequent packets arriving in the secondary bearer are transmitted instantly. The same mechanism can be applied in UL as well, with the RNC sending the end marker and the GGSN buffering the packets in the secondary bearer. This mechanism is transparent to the UE and requires only network-side modification; it was implemented in all simulations presented in this paper.

For the NRSPCA setup latency, two configurations were used in the simulation: it was either set to zero, modeling an ideal case to assess the maximum achievable performance of this technique or it was chosen randomly between 0.8 – 1 seconds at each bearer setup to study the effects of a long and variable setup latency on web browsing performance.

V. SIMULATION RESULTS

The simulated in-bearer prioritization and NRSPCA scenarios are summarized in Table I. The web browsing experience measured by the download rate and latency is shown in Fig. 5, which displays the obtained results for the simulated scenarios at different Iub capacities. Each simulation was executed with five random seeds and the results were averaged

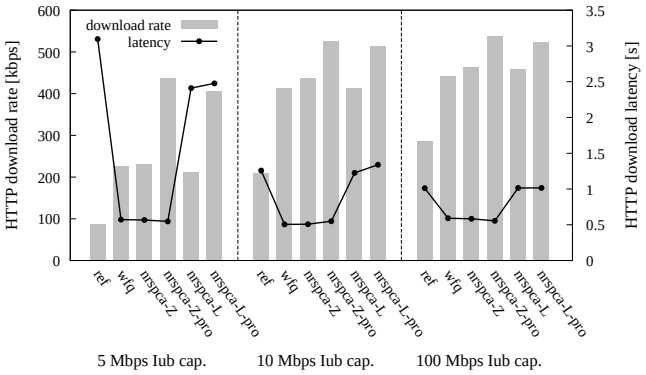


Figure 5. Simulation results showing the average HTTP download rate and download latency at different Iub capacities.

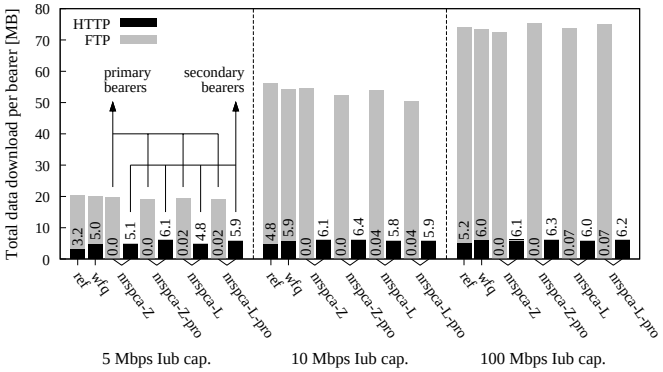


Figure 6. Total amount of data downloaded in the bearers, also visualizing the share between HTTP and FTP. For ref and wfq, there is only one bearer; for the NRSPCA cases, the primary and secondary bearers are both shown. Data labels represent the amount of HTTP traffic in MB.

to obtain the presented data. It is clear that, compared to the reference case, the mechanisms providing the highest HTTP download rate are those involving the setup of a prioritized secondary bearer for HTTP traffic, i.e., scenarios nrspca-Z-pro and nrspca-L-pro (regardless of the NRSPCA setup latency), which is due to the compact end-to-end QoS mechanism provided by the solution; in-bearer prioritization (scenario wfq) also results in considerably high download rate. The impact of the NRSPCA latency on download rate is not significant as despite the latency of setting up the bearer, the vast majority of HTTP traffic is still transmitted in the secondary bearer.

Regarding HTTP download latency, in-bearer HTTP prioritization and immediate secondary bearer setup for HTTP (nrspca-Z or nrspca-Z-pro) considerably reduce the download latency. This is due to that both types of mechanisms prioritize already even the first HTTP packets (either by allocating higher portion of the bandwidth to HTTP by PDCP WFQ or separating HTTP into a secondary bearer by NRSPCA), thereby reducing the queuing delay in the RNC and Node B radio buffers. By comparing nrspca-Z and nrspca-L, it is clear that higher NRSPCA setup latency results in increased HTTP latency that deteriorates user

 ISBN: 978-1-61208-203-5

experience compared to the immediate bearer setup case; the reason is that it is the transmission of the first few HTTP packets that determine the HTTP latency and these packets are still transmitted in the primary bearer without any differentiation until the secondary bearer is established. It should be noted though that the HTTP latency is not worse than the one experienced in the reference case, thus in overall the user experience is improved. Prioritizing the secondary bearer (scenarios nrspca-Z-pro vs. nrspca-Z and nrspca-L-pro vs. nrspca-L) has no significant impact on the download latency as at the beginning of a HTTP session, the underlying TCP connection is still in slow start phase when the main page is requested and sent and there are not many packets on flight that would benefit from an increased wSPI in case of nrspca-Z-pro; also, in case the secondary bearer is set up with a latency, HTTP latency is determined by those packets still transmitted in the primary bearer, which has the same priority in the nrspca-L and nrspca-L-pro cases. Therefore, in case of NRSPCA, it is the separation of the HTTP packets into a secondary bearer and not the promotion of the secondary bearer that principally reduces the radio queuing delay and, consequently, the HTTP latency.

Besides HTTP download rate and latency, the total amount of data transmitted in each bearer was also measured; these results are shown in Fig. 6. In case of in-bearer prioritization, the total amount of data downloaded in the bearer is similar to that of the reference case, with the difference that HTTP represents a higher portion of the overall downloaded data due to the PDCP WFQ mechanism, and due to the fact that the web browsing session was not terminated during the simulation time; better circumstances resulted on increased amount of downloaded pages during the simulation time.

Among all scenarios, the total amount of downloaded HTTP data is the highest if NRSPCA is combined with secondary bearer prioritization (nrspca-Z-pro) since the prioritized secondary bearer does not have to share its bandwidth with other applications, i.e., it is fully allocated to HTTP traffic. With higher NRSPCA setup latency (nrspca-L and nrspca-L-pro), there is also some HTTP data in the primary bearer that is transmitted until the secondary bearer is established; however, the amount is not significant in comparison with the total HTTP data.

VI. Conclusion

In this paper, two alternative application aware mechanisms applicable in HSPA systems, namely the in-bearer prioritization and NRSPCA have been discussed and evaluated based on simulations. The evaluation was focusing on the use case of promoting web browsing traffic over bulk file transfer. Results indicate that NRSPCA is able to separate the applications efficiently. Together with its intrinsic, compact, bearer-based end-to-end QoS mechanisms it provides efficient differentiation and application specific services that outperform the in-bearer mechanisms. The advantage of the in-bearer prioritization is its transparency to the UE, making it a completely network-side solution only requiring support from the RNC; however, the fact that this solution is easily applicable only for DL traffic makes it less attractive. Nevertheless, the advantage of being transparent to the UE makes the in-bearer prioritization a competitive solution compared to NRSPCA-based solutions. Future work in the studied area can be devoted to the analysis of application aware methods in context of additional applications not considered in this paper.

References

[1] H. Holma and A. Toskala, Eds., *HSDPA/HSUPA for UMTS*. John Wiley & Sons, 2006.

[2] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *Network, IEEE*, vol. 24, no. 2, pp. 36–41, Mar. 2010.

[3] D. Soldani, H. X. Jun, and B. Luck, "Strategies for Mobile Broadband Growth: Traffic Segmentation for Better Customer Experience," in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, May 2011, pp. 1–5.

[4] J. Triay, D. Rousseau, C. Cervello-Pastor, and V. Vokkarane, "Dynamic Service-Aware Reservation Framework for Multi-Layer High-Speed Networks," in *Computer Communications and Networks (ICCCN), 2011 Proceedings of 20th International Conference on*, Aug. 2011, pp. 1–7.

[5] H. Holma and A. Toskala, Eds., *WCDMA for UMTS*, 3rd ed. John Wiley & Sons, 2004.

[6] K. Pedersen, P. Mogensen, and T. Kolding, "Overview of QoS options for HSDPA," *Communications Magazine, IEEE*, vol. 44, no. 7, pp. 100–105, Sep. 2006.

[7] D. Laselva, J. Steiner, F. Khokhar, T. Kolding, and J. Wigard, "Optimization of QoS-aware Packet Schedulers in Multi-Service Scenarios over HSDPA," in *Wireless Communication Systems, 2007. ISWCS 2007. 4th International Symposium on*, Oct. 2007, pp. 123–127.

[8] 3GPP, "General Packet Radio Service (GPRS); Service description; Stage 2; R11," 3rd Generation Partnership Project (3GPP), TS 23.060, Dec. 2011.

[9] L. Kőrössy and C. Vulkán, "QoS Aware HSDPA Congestion Control Algorithm," in *Networking and Communications, 2008. WIMOB '08. IEEE International Conference on Wireless and Mobile Computing*, Oct. 2008, pp. 404–409.

[10] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, "Hypertext Transfer Protocol – HTTP/1.1," RFC 2616 (Draft Standard), Internet Engineering Task Force, Jun. 1999, updated by RFCs 2817, 5785, 6266.

[11] "Alexa Top 500 Global Sites," Retrieved Sep. 2011. [Online]. Available: http://www.alexa.com/topsites

[12] 3GPP, "General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U)," 3rd Generation Partnership Project (3GPP), TS 29.281, Dec. 2011.

 ISBN: 978-1-61208-203-5

An ICT-oriented Management Solution for NGNs

Pedro Gonçalves
ESTGA/IT
Universidade de Aveiro
Aveiro, Portugal
pasg@ua.pt

Ricardo Mendes, Rui Aguiar
DETI/IT
Universidade de Aveiro
Aveiro, Portugal
ramm@ua.pt, ruilaa@ua.pt

***Abstract*— NGN architecture reused several standards from the IP world, as exemplified by the Session Initiation Protocol SIP, which is ubiquitous in the majority of these network components. However, the NGN management architecture simply presented a very generic management model that follows TMN. Several management technologies are proposed, such as Web services, CORBA and SNMP, to implement management solutions. Network and systems management standardizing bodies currently promote newer technologies that aim to solve known shortcomings to these. This paper proposes a management solution for NGNs based on recent IP world technologies. The presented solution was implemented in the form of a middleware to manage NGN elements. This middleware was used in the management of an element belonging to the IP Multimedia Subsystem platform, namely the Policy and Charging Rules Function.**

Keywords-component - NGN management; Middleware; technology integration; NETCONF; WBEM.

I. INTRODUCTION

Next Generation Networks (NGNs) started a new era, where the convergence of different worlds truly began: both the convergence of fixed and the merger of telecommunications and the Internet technology become prevalent in NGN. In terms of network convergence, NGN proposed an extremely modular network architecture where control of the transport stratum is agnostic in terms of transport technology, thus allowing to integrate traffic from / to different access networks technologies. As related to technology integration, NGN reused several standards from the IP network world, as exemplified by the Session Initiation Protocol SIP that is ubiquitous in many of these networks components.

Contrary to what happened with the innovative network architecture development, the management architecture specification is quite conventional. The existing documentation specifies a very generic architecture that follows the *Telecommunication Management Forum* (TMF) [1] management model. It proposes several management technologies such as Web services, CORBA and SNMP to implement the management solutions. The standardization bodies that have been working in the definition of management technologies in the area of network and enterprise management are *Internet Engineering Task Force* (IETF) [2] and *Distributed Management Task Force* (DMTF) [3].

In the enterprise management area, the DMTF standardized several technologies like the *Desktop Management Interface* [4], the *Web Based Enterprise Management* (WBEM) [5] and the WS-Management [6], while promoting a vision of integrated management and including the management support for network equipment. In the area of network management, IETF has developed several technologies such as SNMP [7], COPS [8] and more recently NETCONF [9] that addresses several issues raised to the SNMP technology. NGN management has not cope with these evolutions.

This paper proposes a new management approach, exploring the IP-world technologies for the management of NGN. The use of new standards is compatible with the novel TMF management model, but brings the inherent advantages associated with these standards. Given that NGN architecture scope includes aspects related to services and networks, we chose a management approach able to address both areas, relying in technology from system management (WBEM) and a technology from network management (NETCONF). WBEM is an open technology, very flexible, that can easily carry out integrated management of a complex management scenario such as a NGN network. Additionally, and being a popular technology with a vast number of implementations, it enables rapid prototyping of management solutions. NETCONF, on the other hand, has been specified by the IETF to manage network equipment and has been receiving much acceptance by both academy and industry.

In order to allow an integrated management of various aspects of NGN, we proposed a solution that integrates, in the form of an adaptive middleware, the network and systems management. In order to validate our technology adaptation approach, a component from the IMS platform, the Policy Charging and Rules Function (PCRF) [10], was used as a test element. This paper thus describes a management solution that combines the flexibility and comprehensiveness of the WBEM approach with the NETCONF suitability to the computational requirements of network elements. Although our management concept is broader in order to be concise, we focused our implementation on the policy provisioning process, an adapter that allows transfer rules between a central server and a PCRF [10]. Of course, we need to show the reliability of such a mixed approach, witch requires some sort of technology adaptation, able to cope with data model transformation and protocol adaptation.

This paper is organized as follows. Section II gives a general overview of the management technologies involved in this work, and Section III describes related work. Section IV describes the developed system and Section V analyses its performance, and conclusions are finally provided in Section VI.

 ISBN: 978-1-61208-203-5

II. TECHNOLOGY OVERVIEW

WBEM [5] was initially proposed by companies from the desktop management area, and was later developed by the DMTF. WBEM specification includes a set of technologies imported from the web world, such as the HTTP based transport mechanism (*CIM operations over HyperText Transfer Protocol* (HTTP)) [11] and the XML based specification for the information encoding (CIM-XML). The data model used in the WBEM technology is the *Common Information Model* (CIM) [12], a data model proposed by the DMTF that aims to integrate management information of the desktop and of the network areas, witch thus seems specially indicated to our objectives.

CIM is a three-layer object oriented data model, composed of a set of abstract classes and associations that model the generic common characteristics of the management fields such as networks, systems, users, etc., that developers extend in the form of derivate classes. CIM specification describes its basic modeling concepts and meta-schema design; the *Managed Object Format* (MOF) language specifies how information is rendered; and a Schema defines the semantics for a wide range of managed objects and relationships between them. Such a modular and extensible data model allows the integration of multiple management data and enables the development of integrated management solutions. CIM data model was later reused by DMTF in the specification of a new technology based on Web services, named WS-Management.

CIM operations over HTTP [11] specify a vast diversity of operations (named methods in the specification terminology) including methods for classes and qualifier manipulation, methods for instance and property handling, methods for indication dispatch and for class method invocation.

WBEM solutions include four components: the CIM client typically used by the human operator during management tasks, a *CIM Object Manager* (CIMOM) that is the main component of the system maintaining the dialogue with the CIM client and the management information, a CIM repository and CIM providers, that perform the interface between the CIM server and specific managed equipment such as a managed server or a router.

WBEM technology received significant attention by both industry and academy, having been used as the enabling technology management solutions for various scenarios [13-16]. Despite the completeness of its data model, WBEM technology is mainly used in the area of system management.

WBEM architecture requires that providers be created for handling the CIM extensions. Given that, providers act as an adapter between the management server and the managed elements. Providers are the most appropriate element in the architecture to implement adaptation to management technologies. This can explain the associate amount of work in the literature [14, 15, 17, 18]. In [15], Yoon et al. describe a WBEM-based management system for residential gateways that interfaces the managed equipment through a SNMP interface. System features include equipment configuration, performance monitoring and equipment fault detection. Seo et al. [14], describe a *Network Management System* (NMS) for managing DiffServ-over-MPLS QoS in an inter-domain scenario, where in addition to interconnect SNMP and WBEM, it performs admission control using a COPS-RSVP interface, implemented in a dedicated provider.

On the other hand NETCONF [9] is a management protocol standardized by the IETF in 2006 that defines operations for managing network devices, allowing to upload, retrieve and manipulate management configuration data. The protocol is based on a XML-encoded Remote Procedure Call (XML-RPC) over secure transports as SSH [19], SOAP [20], TLS [21] or BEEP[22]. In order to maintain the interoperability between NETCONF management solutions it was decided that SSH transport implementation was mandatory. The protocol modularity was promoted by means of a architecture composed of four layers: the content layer containing configuration data; the operations layer implementing the management operations; the RPC layer implementing the XML-RPC remote procedure call; and the transport protocol layer that implements the information transport between management entities. Although initial specification just included configuration operations, it was later standardized the NETCONF monitoring support [23].

The protocol was designed independently of the management data model, and therefore the RFC that specify the protocol does not include any considerations about the data model. So, the IETF NETMOD working group proposed a data modeling language named YANG [24] for the generation of the management data models. YANG defines a set of data nodes organized as a hierarchical tree. Each data node in the tree has a name and either a value or a set of child nodes. YANG schema is structured into modules and sub-modules that may be published by a standards organization, an enterprise or an industry forum [25]. A key YANG design feature is the modular extensibility. One YANG module may define additional data nodes augmenting the data nodes defined in another YANG module. YANG was adopted in this work as the NETCONF data model language.

Although additional operations could be provided based on the capabilities advertised by the devices, the base protocol defines nine operations [9] for datastore data management and two operations [23] for the management of event notification information.

During the recent years NETCONF has received much attention by both academia and industry, and several NETCONF *Software Development Kits* were developed [26-29].

Ozianyi et al. [13] describe an XML-driven framework for Policy-based QoS management for a IMS network. The proposal consists in a NETCONF *Network Management System* (NMS) that implements policy based management of IMS *Policy Control and Charging* subset [10], integrated with a variety of network elements that include COPS-PR, Diameter and NETCONF support. They assess the bandwidth utilization and the communication delay between the management elements of the NETCONF and the COPS-PR technologies, concluding that compressed NETCONF interfaces perform better than COPS-PR interfaces, for some higher values of information (transferred configuration composed of more policies).

 ISBN: 978-1-61208-203-5

More recently Enns et al. proposed an update [30] to RFC 4741 where they add a YANG module for NETCONF operations, removing his description from the XML Schema Definition. Also they create a username and the requirement for NETCONF servers to perform user authentication and permissions authorization according to the user profile. It has been developed by Perelman et al. [31] a reduced version of NETCONF original protocol, named *NETCONF Light*, that includes a subset of the original protocol, envisioned for devices with limited computing resources. Among the main differences from the original protocol highlights the lack of support filtering configuration for *NETCONF light* operations. Despite maintaining original protocol operations, this light version removes the possibility of defining a filter that limits the operations scope over the equipment configuration.

III. SYSTEM DEVELOPED

In our management concept, CIM server acts as the central management point of the IMS platform providing all the flexibility required for an integrated service and network management we then resort to NETCONF to network management, for flexibility and efficiency [32]. A new middleware then performs the technology adaptation between WBEM and the devices running NETCONF. Figure 1 illustrates the overall system architecture from the management application to the managed device.

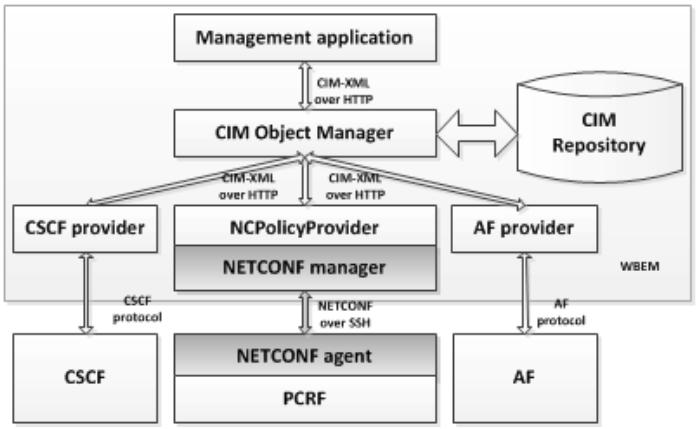


Figure 1. System architecture

As an implementation use case, we centered in the management of the PCRF, an IMS functional entity that performs admission control, resource management and charging, based on policies. For simplicity of discussion, we focused here on the development of a middleware that allows that policies specified by the system administrator in the NMS, to be pushed via NETCONF over SSH interface to the PCRF. In this section we describe the information model and the message sequences between the CIM server and the NETCONF agent.

The NMS has been developed based on a CIM server with some extensions to the CIM data model. It was used a CIM client where the administrator performs the specification of policies, which in turn are delivered to the NMS. The system includes a repository that maintains the previously defined policies. A CIM provider that implements a NETCONF interface and communicates with the NETCONF agent makes the adaptation of policy information. The developed middleware consisted in an *OpenPegasus* provider based on a development platform named CIMPLE [33] and was developed in the C language. The development followed a modular architecture, having been created logic to decode WBEM messages, to encode XML components and to implement a NETCONF manager. The XML messages encoding is performed using a *Xerxes-C* library and the implementation of the NETCONF manager has been carried out using a NETCONF over SSH implementation [27].

A. Data model

Policies that are used by the system define the behavior of PCRF, determining how to make admission control and how to manage network resources. A system policy is formally defined as an aggregation of policy rules. Each policy rule is composed of a set of conditions and a corresponding set of actions. The condition defines when the policy rule is applicable. Once a policy rule is activated, one or more actions contained by that policy rule may be executed. In our middleware development of reused some previous work [34] as well as the policy representation data model.

The data model associated with NETCONF component was derived from the CIM component data model. CIM classes were converted into *leaflists* and association classes were converted as *leafrefs*. CIM class properties were implemented as leafs of a YANG equivalent built-in data type and were placed as leaf elements inside the *leaflists*. CIM class methods were declared as NETCONF operations with YANG module scope, since the language does not allow the method declaration with a scope associated with YANG constructs.

B. Operation encoding

Technology adaptation implies, besides the data model conversion, the matching between the operations of both technologies. TABLE 1 identifies the immediate matching between the WBEM and NETCONF operations.

TABLE 1. WBEM and NETCONF operations match

WBEM	NETCONF
GetInstance	<get-config>/<get>
EnumerateInstances	<get-config>
CreateInstance	<edit-config>
ModifyInstance	<edit-config>
DeleteInstance	<delete-config>
ExportIndication	<notification>

WBEM operations have a granularity of the object, given the nature of its object oriented data model. When some WBEM operation is performed, it affects an object. In the case of NETCONF, when some operation is performed it affects a complete document or a part of it, if a filter expression is provided to the operation. As formerly described, policies are represented as the aggregation of objects from several classes, and they are transferred in WBEM technology in a per-object basis.

The message match mechanism described in the TABLE 1 is extremely inefficient, since it would execute a NETCONF operation sending a small policy component, when it could

 ISBN: 978-1-61208-203-5

receive the complete WBEM policy description and then send the full policy to NETCONF agent with a single operation. For instance in the *read policy* use case, the system needs several *GetInstances* to visualize the entire policy but only one <*get-config*> to obtain the total policy from the managed device. On other hand, depending on the NETCONF agent capabilities, the <*edit-config*> should be preceded by the <*lock*> operation and succeeded by the <*commit*> and <*unlock*> operation, to atomize the submit operation. Most cases follow this trend: the direct operation match is inadequate. For example the policy creation process requires several WBEM *CreateInstance* operations and a single atomic <*edit-config*> operation in the NETCONF technology. Figure 2 illustrates the create policy use case where our middleware receives the complete list of policy components from WBEM component, and after recoding it, performs an atomic policy transfer to the NETCONF agent.

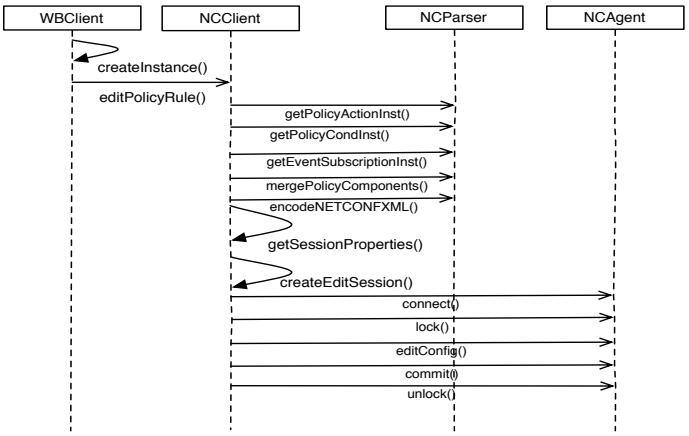


Figure 2. Policy creation sequence diagram

IV. SYSTEM ANALYSIS

This section presents the scenario used to evaluate the management system, defines the evaluation uses cases and evaluates the results.

A. Test scenario

The test scenario included three machines running the different applications: a machine that ran the CIM client, another that ran the CIM server with the developed provider and a third machine with the NETCONF agent. The network segment was isolated from the local LAN in order to avoid the network traffic to bias the evaluation results. The WBEM server was installed in a 2.0 GHz Intel Centrino with 1GB RAM, and the client was installed in a 2.6 GHz Pentium 4 core with 512 MB RAM, both running 2.6.34 Linux Kernel.

The following use cases, typical for a PCRF, were considered for system analysis: (i) read the running policy from the Managed Device; (ii) edit the property "*Caption*" from the "*UA_ActionNull*" class and submit the changes; (iii) create a policy from scratch. For each use case the tests were performed, the traffic was captured and analyzed on the WBEM and on NETCONF interfaces of the middleware, and the memory consumption for each measured.

B. Traffic analysis

Upon execution of the three test cases, the WBEM network traffic captures were filtered, analyzed and quantified. TABLE 2 contains for each test, the amount of traffic discriminated by each of the layers of the protocol stack, as well as their relative weight. To simplify the traffic analysis, the Secure Sockets Layer (SSL) support was separated of the rest: the session establishment process includes the connection and the authentication of the client in the WBEM server.

TABLE 2. WBEM traffic

	Session Establish.		Read Policy		Edit Policy		Create Policy	
Packets	8		417		10		601	
Messages	1		102		2		45	
Comp.	Kb	%	Kb	%	Kb	%	Kb	%
CIM-XML	0	0	206,3	69.68	4,2	73.07	103,9	61.33
HTTP	0,8	60.51	62,9	21.24	0,9	15.59	26,8	15.8
TCP	0,3	19.15	13	4.4	0,3	5.5	18,8	11.09
IP	0,2	11.97	8,1	2.75	0,2	3.43	11,7	6.93
Ethernet	0,1	8.38	5,7	1.93	0,1	2.4	8,2	4.85
Total	1,3	100	296	100	5,7	100	169,4	100

A *read policy* operation requires that 417 packets and 102 messages are exchanged between CIM entities until the complete policy could be presented in the CIM client. A deeper analysis in the traffic exchanged shows that before showing any class, the CIM client performs at least three requests to the CIM server: *GetClass*, *EnumerateInstanceNames* and *GetInstance*. Additionally it was observed that the CIM client repeated those operations for each instance it receives, thus requiring a high number of operations. Apart from the messages exchanged between the CIM client and server, the generated traffic is further increased by the HTTP transport overhead, which exceeds 20% of generated traffic.

By its turn, the *edit policy* is less demanding on traffic, just 10 packets and 2 exchanged messages. In this case the high fragmentation of the policy in separated classes may be an advantage, because we only need to retrieve the policy class that contains the property that is intended to edit. For the policy creation, it took 601 packets and 45 messages. Paradoxically this test case requires more packets and less information (more than half on the read case) despite being the same policy. In this case, there are less messages exchanged, 45 versus 102. Editing a class requires two requests to the CIM server: *GetClass* and *CreateInstance*.

Upon execution of the three test cases, the NETCONF network traffic captures were filtered, analyzed and quantified. TABLE 3 contains for each test, the amount of traffic discriminated by each component as well as their relative weight. We further indicate the load of session establishment. As before the session establishment corresponds to the establishment of an SSH connection with its key exchange between the NETCONF Manager and NETCONF Agent.

The *read policy* use case from the NETCONF component, involve obtaining the policy from the running repository through the operation <*get-config*>. This operation is quite

 ISBN: 978-1-61208-203-5

simple, requiring only one message. The policy itself represents more than 80% of the traffic exchanged. For the edit policy use case, the changes are submitted by the NETCONF Manager, sending the policy component containing the change. The system policy follows the conceptual approach, events, conditions and actions: <*UA_EventSubscription*>, <*UA_PolicyCondition*> and <*UA_PolicyAction*> respectively. In this case exists a bigger granularity in the submitted changes comparing with WBEM, and by other way the NETCONF agent supports the candidate capability which means that four requests must be made to submit the changes in an atomic way: <*lock*>, <*edit-config*>, <*commit*> and <*unlock*>. Finally the create policy case, is the operation that spends more traffic, for the same reasons pointed before in the edit policy case. Entire policy plus the <*lock*>, <*edit-config*>, <*commit*> and <*unlock*> requests.

TABLE 3. NETCONF traffic

	Session Establish.		Read Policy		Edit Policy		Create Policy	
Packets	41		23		19		27	
Messages	4		1		4		4	
Comp.	Kb	%	Kb	%	Kb	%	Kb	%
XML	0	0	15,8	80.17	8,8	66.78	18,3	76.97
RPC	0	0	0,2	1.2	0,9	7.14	0,9	3.99
SSH	7,3	73.92	2,2	11.12	2,2	16.82	2,8	11.7
TCP	1,3	12.93	0,7	3.65	0,6	4.49	0,8	3.56
IP	0,8	8.08	0,4	2.28	0,4	2.8	0,5	2.22
Ethernet	0,6	5.66	0,3	1.6	0,3	1.96	0,4	1.56
Total	9,9	100	19,7	100	13,2	100	23,7	100

C. *Memory requirements*

An analysis was performed to the memory consumption in the WBEM server, in the SSH daemon and in the NETCONF agent during an entire edit operation, which comprehends the session establishment, the read and the edit policy procedures. The SSH daemon and the NETCONF agent showed a constant behavior during the entire process: SSH daemon occupied 580kB and the NETCONF agent 1212kB. The memory used by the WBEM server grew from operation to operation: 8924kB when the service was started, 9456kB when the session was established, 11076kB when *NetConfSession* instance was created, 14152kB when the policy information was read and 15052kB when the policy information was edited.

The software was subjected to profiling tools in order to detect problems with the resource usage (memory, processing). It analyzed memory usage; the time spent executing each method and the sequence of the application methods invocation. A framework named *Valgrind* was used to detect errors resulting from incorrect use of dynamic memory and a helper application named *Callgrind* recorded the call history among functions in a program's run, as a call-graph. The profiling analysis made to the system showed that there are no problems in memory usage and confirmed that results are coherent, since both emphasize *Xerces-C* methods as the most time and memory consuming. This is easily justified, since *Xerces-C* is used for parsing XML, and this corresponds to the major part of the processing.

D. *Load analysis and profiling analysis*

They were some load tests run in order to verify system stability and scalability. Tests were done using a command line CIM client (*cimcli*), invoked by some bash scripts that perform read policy requests to the system. Two types of tests were performed: a test where the WBEM server made several calls to different agents conducting a NETCONF request, and another test with the WBEM server making several requests to a same agent.

Figure 3 illustrates the memory behavior of the WBEM server as result of the multi agent scenario. Tests show an increase in memory consumption, which means that the limitation for the reception of new requests from the CIM server is the memory of the machine where the process is running. Indeed this is what happens, because of 3500 applications, the CIM server was only able to process 2961, reaching the limit of available memory (2GB), and the server process was killed by the O.S.

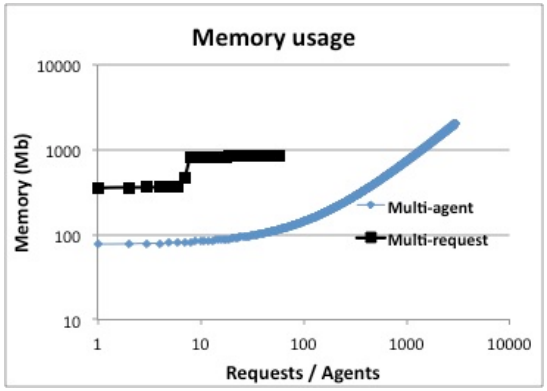


Figure 3. Memory behavior in load tests

In the single agent scenario, tests performed various requests to a same agent; the requests were executed in parallel, imposing a much higher load to the server. Figure 3 illustrates the behavior of memory consumption in this situation. In this case the memory behaves differently due the simultaneous requests and does not constitute the limitation because the limit has not been reached. However the processing power in server capacity CIM to answer the multiple simultaneous requests in a timely manner. The system was able to process 58 requests, rejecting the others by a timeout.

V. CONCLUSION AND FINAL DISCUSSION

This paper presents a NGN management solution based on ICT-world technologies. It consists in a CIM server middleware that integrates the two management technologies from the enterprise and the network management. WBEM is a widely used technology in the system management area, with plenty of implementations and commercial products. NETCONF is a newer network management technology that addresses SNMP shortcomings and has been receiving lots of attention.

A management solution consisting of a technology integrating middleware allows by on hand to implement a management system capable of managing a vast diversity

 ISBN: 978-1-61208-203-5

equipment, as the functional entities that exist within the IMS platform; and by the other hand it allows to adapt the technologies, using the most appropriate management technology for each equipment type. Additionally, and using an integration technique that integrates the data models used in the telecom and the enterprise management [35], the proposed solution could cope with upper layers of the telecom management technology.

Although a broader architecture was proposed, the discussion was centered on the implementation of the policy provisioning process, and the developed middleware was applied to the management of an IMS element named PCRF. The middleware was functionally evaluated and stress tests conducted in order assess its scalability and its applicability to a production scenario.

The results show that these technologies scale in a promising way to its usage in a production network and, especially the need for NETCONF was evident considerably more efficient than WBEM closer to network elements. WBEM technology offers greater flexibility, allowing the management of a greater equipment range, given the vast richness of its data model.

REFERENCES

[1] TM Forum, http://www.tmforum.org/browse.aspx, 2012-04-09.

[2] The Internet Engineering Task Force (IETF), http://www.ietf.org/, 2012-04-08.

[3] DMTF, Desktop Management Task Force Home, http://www.dmtf.org/home, 2012-04-09.

[4] DMTF, Desktop Management Interface Specification - v 2.0.1, v 2.0.1, 2003.

[5] J. P. Thompson, "Web-based enterprise management architecture", Communications Magazine, IEEE, vol. 36, pp. 80-86, 1998.

[6] DMTF, Web Services for Management (WS-Management), 2005.

[7] J. Case, M. Fedor, M. Schoffstall, and J. Davin, Simple Network Management Protocol (SNMP), RFC 1157, 1990.

[8] D. Durham, J. Boyle, R. Cohen, S. Herzog, R. Rajan, and A. Sastry, The COPS (Common Open Policy Service) Protocol, RFC 2748, 2000.

[9] R. Enns, NETCONF Configuration Protocol, RFC 4741, 2006.

[10] J. J. P. Balbas, S. Rommer, and J. Stenfelt, "Policy and charging control in the evolved packet system", Communications Magazine, IEEE, vol. 47, pp. 68-74, 2009.

[11] DMTF, Specification for CIM operations over HTTP version 1.2, 2007.

[12] DMTF, "Common Information Model (CIM) Specification - Version 2.28", 2011.

[13] V. G. Ozianyi, R. Good, N. Carrilho, and N. Ventura, "XML-Driven Framework for Policy-Based QoS Management of IMS Networks", Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE, 2008, pp. 1-6.

[14] J.-C. Seo, H.-S. Kim, D.-S. Yun, and Y.-T. Kim, "WBEM-Based SLA Management Across Multi-domain Networks for QoS-Guaranteed DiffServ-over-MPLS Provisioning", in Management of Convergence Networks and Services. vol. 4238, ed: Springer Berlin / Heidelberg, 2006, pp. 312-321.

[15] B. W. Yoon, S. Ahn, and J. W. Chung, "Web-Based Home Gateway Management System using SNMP", 17th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management Large Scale Management, Dublin, Ireland, 2006.

[16] J.-h. Yoon, H.-t. Ju, and J. Hong, "Development of SNMP-XML translator and gateway for XML-based integrated network management", Int. J. Netw. Manag., vol. 13, pp. 259-276, 2003.

[17] S.-J. Lee, M.-J. Choi, S.-M. Yoo, J. W. Hong, H.-N. Cho, C.-W. Ahn, and S.-I. Jung, "Design of a WBEM-based Management System for Ubiquitous Computing Servers"

[18] Y.-J. Oh, H.-T. Ju, M.-J. Choi, and J. W.-K. Hong, "Interaction Translation Methods for XML/SNMP Gateway", Proceedings of the 13th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management: Management Technologies for E-Commerce and E-Business Applications, 2002.

[19] M. Wasserman and T. Goddard, Using the NETCONF Configuration Protocol over Secure Shell (SSH), RFC 4742, 2006.

[20] T. Goddard, Using NETCONF over the Simple Object Access Protocol (SOAP), RFC 4743, 2006.

[21] M. Badra, NETCONF over Transport Layer Security (TLS), RFC 5539, 2006.

[22] E. Lear and K. Crozier, Using the NETCONF Protocol over the Blocks Extensible Exchange Protocol (BEEP), RFC 4744, 2006.

[23] S. Chisholm and H. Trevino, NETCONF Event Notifications, RFC 5277, 2008.

[24] M. Bjorklund, YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF), RFC 6020, 2008.

[25] E. Nataf and O. Festor, "End-to-end YANG-based configuration management", Network Operations and Management Symposium (NOMS), 2010 IEEE, 2010, pp. 674-684.

[26] B. Zores, R. State, and O. Festor, YENCA, http://sourceforge.net/projects/yenca/, 2011-03-22.

[27] Netopeer - Remore configuration system using NETCONF protocol, http://code.google.com/p/netopeer/, 2012-04-21.

[28] Yuma - YANG-based Unified Modular Automation Tools, http://sourceforge.net/projects/yuma/, 2012-04-21.

[29] P. Tavares, P. Gonçalves, and J. L. Oliveira, "An IDE for NETCONF management applications", 7th Latin American Network Operations and Management Symposium, Quito, Equador, 2011.

[30] R. Enns, M. Bjorklund, J. Schoenwaelder, and A. Bierman, Network Configuration Protocol (NETCONF), RFC 6241, 2011.

[31] V. Perelman, J. Schoenwaelder, and M. Ersue, Network Configuration Protocol for Constrained Devices (NETCONF Light), draft-schoenw-netconf-light-00.txt, 2011.

[32] P. Gonçalves, J. L. Oliveira, and R. L. Aguiar, "An evaluation of network management protocols", 11th IFIP/IEEE International Symposium on Integrated Network Management (IM 2009), New York, USA, 2009.

[33] M. E. Brasher and K. Schopmeyer, "CIMPLE: An Embeddable CIM Provider Engine", March 2006.

[34] P. Gonçalves, R. Azevedo, J. L. Oliveira, and R. Aguiar, "Managing QoS in a NGN using a PBM approach", The Fourth International Conference on Systems and Networks Communications- ICSNC 2009 Porto, Portugal, 2009.

[35] TMForum, CIM-SID Solution Suite Release 6.0, 2006.

 ISBN: 978-1-61208-203-5

Centralized Bandwidth Management in Multi-Radio Access Networks

Balázs Héder
Nokia Siemens Networks
Budapest, Hungary
balazs.heder@nsn.com

Péter Szilágyi
Nokia Siemens Networks
Budapest, Hungary
peter.1.szilagyi@nsn.com

Csaba Vulkán
Nokia Siemens Networks
Budapest, Hungary
csaba.vulkan@nsn.com

***Abstract*—Radio access technology evolution resulted in two alternative architectural solutions: Evolved HSPA (High Speed Packet Access) systems with centralized architecture and LTE (Long Term Evolution) systems with distributed, full packet based architecture. Both systems are capable of providing high data rates and low latency to the users. Due to factors such as the need to preserve existing investments and reduced operational costs, for the time being these systems will coexist by sharing a common transport infrastructure and by providing services over the same areas. Good user experience over these systems requires harmonized QoS (Quality of Service) architectures and fair resource sharing mechanisms even in case of transport congestion. Technological and architectural differences of HSPA and LTE systems result in fairness problems that are not handled well by existing mechanisms designed for homogeneous environments. This paper proposes a comprehensive solution which, as simulation results indicate, has superior performance and handles the fairness and QoS issues efficiently.**

***Keywords*-HSPA, LTE, CC, multi-RAN, QoS**

I. INTRODUCTION

Smart phones are able to provide true multimedia experience and access to the multitude of Internet based applications and services such as streaming multimedia, mobile mail, web browsing, instant messaging, micro blogging, etc., which dominantly use TCP (Transmission Control Protocol) as transport protocol. This generates continuously growing demand for increased radio access system capacity, high user data rates and reduced latency. In parallel with the penetration of smart devices, the radio access technology is evolving as well. There are two main tracks of this evolution defined by the 3GPP (3rd Generation Partnership Project): evolved HSPA and LTE. On the one hand, evolved HSPA improves the radio and transport capability of the WCDMA (Wideband Code Division Multiple Access) systems via additional functionalities mainly implemented at the Node B without changing the system architecture. On the other hand, LTE proposes a full packet based technology with new, flat architecture where the radio and the transport network layers are packet switched and radio protocols are terminated at the eNBs (evolved Node Bs). In LTE, the latency of packet transmission is low because there are no Radio Layer 2 RTXs (retransmissions) over the transport network as opposed to the WCDMA/HSPA. Existing radio access networks based on WCDMA/HSPA technology will not necessarily be replaced by LTE but will coexist with it in a heterogeneous environment, where in certain locations multiple radio access possibilities (WCDMA, HSPA, LTE, etc.) will be provided to the users. This coexistence increases the system capacity and diversity, preserves the existing investments and provides a fall-back possibility and redundancy. As the LTE transport network layer is already packet based and HSPA is being migrated over packet technology, the deployment of a common transport network to be shared by the coexisting radio access systems is an obvious choice that allows efficient management and resource usage. These heterogeneous systems are referred to as multi-RANs (Multi-Radio Access Networks) in this paper. Harmonized QoS over multi-RANs is an important enabler of proper user experience. Users should have the same experience regardless of their point of attachment, that is, they should be able to use their applications with acceptable quality both over HSPA and LTE. Harmonized QoS has two important enablers: consistent HSPA and LTE QoS parameters, and QoS enforcement mechanisms able to provide fair resource usage over the shared transport. The former means that HSPA and LTE UP (user plane) bearers providing the same service should have a set of compatible QoS parameters. The latter requires coherent mapping to transport services. Assuming packet transport with DiffServ (Differentiated Services) based QoS architecture, this can be achieved by marking packets of the same application/service with the same DSCP (DiffServ Code Point) regardless of the access technology (HSPA or LTE). While the definition of harmonized HSPA and LTE QoS parameters and mapping rules is a simple management task for operators, QoS enforcement also raises problems that are not of administrative nature. Transport congestion that might occur in packet based networks (especially on the capacity limited backhaul links such as microwave radio) is handled differently in legacy (HSPA) and flat (LTE) systems. This is due to the difference in architecture and to technological constraints, such as the operation of the Radio Layer 2 protocols in HSPA systems. The HSPA CC (congestion control) mechanism, introduced by 3GPP [1], has the additional scope to prevent RLC AM (Radio Link Control Acknowledged Mode) RTXs over the Iub interface

 ISBN: 978-1-61208-203-5

[2] as these can cause significant efficiency degradation. LTE has no such standardized solution; currently, it relies on the TCP CC mechanism, that, together with RED (Random Early Detection), is able to resolve congestion and enforce fairness among the connections. In LTE, this might be enough but not for HSPA as it is not able to prevent RLC AM RTXs [3].

When the transport is shared by the LTE and HSPA traffic, congestion may cause fairness problems as HSPA traffic is not TCP friendly, i.e, the TCP sources can achieve only a limited throughput when competing for transport resources with TCP unfriendly traffic [4].

The coexistence of GSM (Global System for Mobile Communications), WCDMA and LTE on a shared transport is mentioned in [5] but it does not deal with the fairness problems in multi-RAN. An idea to use TCP friendly rate control in HSDPA (High Speed Downlink Packet Access) is described in [6] but considering only a homogeneous environment. An alternative HSDPA CC algorithm based on PDCP (Packet Data Convergence Protocol) / RLC packet discard was presented in [7] that solves the fairness problem only in case of DL congestion. Also, the applicability of the solution is limited to TCP.

This paper discusses the problems of inter-system fairness over capacity limited transport networks shared by multi-RAN systems. A novel centralized CC and bandwidth management algorithm is proposed, capable of resolving congestion and enforcing the right level of QoS and fairness. TCP and UDP (User Datagram Protocol) based user traffic are handled in the same way, without compromising the QoS and fairness. The solution is flexible, i.e., it can be used both in homogeneous and heterogeneous systems. The actions of the CC are based on the actual status of the system, the available resources, the topology, the QoS and fairness policies.

The rest of the paper is organized as follows. Section II provides a detailed overview of the multi-RAN systems, defines the fairness criteria and QoS requirements and deals with the fairness problem in case of transport congestion. Section III describes the proposed centralized CC algorithm. Performance evaluation is given in Section IV and finally Section V concludes the paper.

II. System Overview

Multi-RAN systems are based on the cooperation of the HSPA and LTE network elements. HSPA and LTE specific architectural elements impose special fairness and QoS aspects whereas transport congestion requires a common CC.

A. The System Architecture of Multi-RAN Systems

The architecture of a multi-RAN system [8] (Fig. 1) consists of HSPA and LTE network elements connected by user and control plane interfaces. Access to the packet services is granted through the SAE-GW (System Architecture Evolution Gateway). The eNBs are connected directly to the SAE-GW via the S1-U interface. HSPA traffic can reach the CN (core network) through the Iub that connects the Node Bs to the RNC (Radio Network Controller). The RNC is connected to the SGSN (Serving GPRS Support Node) via the Iu-PS interface. The S1-C and S11 interfaces provide the LTE control plane connectivity. The MME (Mobility Management Entity) is responsible for the UE authentication, location tracking and subscription profile management within the LTE system. Inter-system control plane connectivity is available via the S3 interface, whereas the S4 interface provides mobility and control support between the SGSN and the SAE-GW. From the RNC's point of view, the SAE-GW takes the role of the GGSN (Gateway GPRS Support Node). The RNC is connected to the SAE-GW via the S12 interface when direct tunnel is established and indirectly via the Iu-PS and S4 interfaces when no direct tunnel is established. The S12 is based on the Gn-u interface between the SGSN and GGSN in the legacy architecture (not shown).

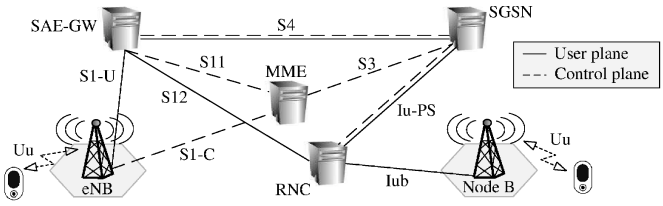


Figure 1. System architecture of a heterogeneous multi-RAN system

B. Harmonized QoS in Multi-RAN Systems

The HSPA and LTE QoS architectures are bearer centric, that is, the QoS parameters are defined and enforced on bearer level. HSPA bearers and LTE EPS (Evolved Packet System) bearers responsible for the UP connectivity between the CN and the UE are mapped to RABs (Radio Access Bearers) by the Radio Network Layer protocols terminated at the RNC (HSPA) and at the eNB (LTE), respectively. In both systems, the air interface packet scheduler has key role in the QoS enforcement, therefore at bearer/RAB setup, the related QoS parameters are signaled to the Node B/eNB. The Node B receives the RAB specific QoS parameters through the RNC: the SPI (scheduling priority indicator), the GBR (guaranteed bit rate) and the DT (discard timer) [9]. The SPI allows the definition of at most 16 distinct priorities. For each SPI, a GBR and DT value can be defined. HSPA flow and congestion control mechanisms support the packet scheduler in the QoS enforcement. The LTE systems allow the definition of 9 distinct QoS classes, referred to as QCI (Quality Class Identifier) classes, that is, upon setup, each EPS data bearer and the corresponding LTE RAB are mapped to a QCI [10]. For each QCI, and thus for each bearer, a GBR value can also be defined. At the transport network, HSPA bearers, RABs and EPS bearers are mapped to the transport QoS classes by DSCP

 ISBN: 978-1-61208-203-5

marking. For each SPI or QCI, a separate DSCP can be used. Note that the transport network QoS architecture should be configured so that it gives full support to the HSPA or LTE QoS. These parameters and mechanisms are sufficient for QoS enforcement in homogeneous radio access systems. Fairness is achieved if at a given Node B or eNB, bearers having the same SPI or QCI respectively receive the same level of service whereas bearers having different SPI or QCI receive service proportional to their QoS parameters. First, the packet scheduler should enforce the GBR of the bearers, whereas the remaining air interface resource should be distributed by considering the priority of the bearers. Throughout this paper, we assume that both the HSPA and LTE air interface packet schedulers implement the PF-RAD (Proportional Fair with Required Activity Detection) discipline [11], which is able to achieve optimal air interface usage and QoS differentiation. In order to facilitate the relative prioritization of the bearers, for each SPI/QCI an additional parameter, the scheduling weight (*wSPI* and *wQCI* respectively) is configured at each Node B/eNB. For the sake of simplicity and without loss of generality, we assume in this paper that the GBR of the bearers is zero, that is, QoS differentiation is enforced solely based on the *wSPI* and *wQCI* parameters. Fairness and QoS differentiation between the QoS classes i and j is achieved if the following expression is true: $\tau_i/\tau_j \approx w_i/w_j$, where τ_i and w_i denote the average measured throughput and the weight of QoS class i, i.e., the *wSPI* in case of the HSPA and the *wQCI* in case of the LTE. In multi-RAN systems, not only the intra- but the inter-system fairness must be achieved as well, i.e., user traffic belonging to the same application should receive the same relative service both through HSPA and LTE. One possibility is to give global meaning to the system specific QoS parameters, i.e., within the multi-RAN system, common QoS classes are defined with a set of well defined common data bearer and RAB level QoS parameters (GBR, weight, etc.). HSPA and LTE bearers are mapped to these classes and their own parameters are derived from these common QoS parameters. The inter-system fairness criteria is that $\tau_i/\tau_j \approx w_i/w_j$, $\forall i,j \in$ HSPA or LTE bearer, that is, the inter-system fairness is met if τ_i/w_i (the measured and weighted average throughput) is approximately the same for each QoS class in each radio access technology. In this setup, there is no need for dedicated bandwidth allocation to HSPA or LTE traffic over the transport network, thus the transport network is truly a shared resource, allowing the maximization of the multiplexing gain. That is, the resources can be dynamically shared by the HSPA and EPS bearers.

C. The Impact of Transport Congestion

In heterogeneous systems, LTE and HSPA share the same transport network as deploying separate transport for each RAN is not a realistic option due to cost, efficiency and manageability reasons. Despite the capabilities of the backhaul transport protocols (resilience, high data rate, low latency, QoS differentiation, etc.), transient congestion may occur due to the capacity limited links such as microwave radio or due to the overbooking of the high capacity aggregation links. During congestion, connections experience increased delay, packet drops and reduced throughput; additionally, it may deteriorate the intra- and inter-system fairness as well. Therefore, efficient CC mechanisms are needed. TCP, the dominant transport protocol used by the majority of data applications, has its own CC mechanism that reacts to congestion by reducing the rate of the connection and by retransmitting the data that is assumed to be lost. Together with RED, it is able to enforce fairness as well. In flat systems such as LTE, where packet drops due to transport congestion are transparent to the Radio Network Layer protocols, TCP's end-to-end CC mechanism is sufficient provided that its latency or the experienced RTT (Round Trip Time) is acceptable. In contrast, packet drops on the transport links connecting the Node Bs to the RNC trigger RLC AM RTX that has negative impact on the overall HSPA performance. The functionality of the HSPA systems has been extended by 3GPP [1] with means of detecting congestion without specifying the CC algorithm itself. The specified framework reuses the existing features of the HSPA systems and, despite the technical differences, provides similar solutions for UL (HSUPA, High Speed Uplink Packet Access) and DL (HSDPA). The HSPA CCE (CC Entity) is located at the Node B and it controls the rate of the connections either via capacity allocations sent to the RNC (HSDPA) or via grants issued to the UEs (HSUPA). Congestion detection is possibly based on the Delay Reference Time and Sequence Number IEs (Information Elements) included in the HS-DSCH (High Speed Downlink Shared Channel) and E-DCH (Enhanced Dedicated Channel) FP (Frame Protocol) data frame headers. The information provided by these IEs are used to detect delay build up (a common solution is to compare the estimated delay against thresholds) or packet drop (as frames are delivered in sequence, a missing sequence number indicates a drop).

In DL, congestion is detected at the Node B [3], [12], whereas UL congestion is detected at the RNC that informs the Node B about it through the E-DCH FP CI (Congestion Indication) control frame messages [13]. The CCE at the Node B reacts to the detected DL congestion by reducing the resource grants of the flows via Capacity Allocation messages sent to the RNC. In a similar way, upon the reception of the CI, the Node B reduces the UL air interface resource grants to be provided to the UEs.

Efficient HSPA CC algorithms are not only being able to resolve transport congestion but can also support the HSPA QoS architecture by considering the QoS parameters of the active bearers at CC decisions. The delay measurement is an important element of the HSPA CC: delay must be kept low so that random discards by RED are avoided and

 ISBN: 978-1-61208-203-5

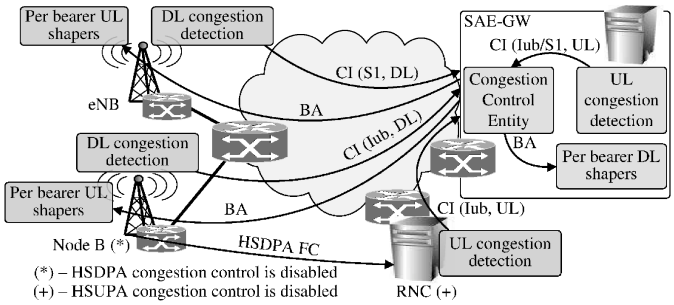


Figure 2. Concept of the centralized congestion control

(or) RLC timer expiration is prevented, i.e., the CC should keep the transport buffers under moderate load in order to prevent RLC AM RTXs. For further details on HSPA CC implementation, the readers are referred to [3].

As for the reasons discussed above and because the Node B and the RNC are topologically closer to each other than the UE and the content servers, the HSPA CC feedback loop is shorter than the end-to-end TCP CC loop. Therefore, in case the narrow link is shared by HSPA and LTE, the rate of the HSPA bearers is reduced first upon congestion. The unused bandwidth is taken by TCP connections over LTE, which continues until the total starvation of the HSPA bearers [7]. Disabling the HSPA CC in multi-RAN environments in order to prevent the self-starvation of HSPA bearers is not a good option either as at congestion, RLC AM RTXs over the Iub cause not only HSPA performance degradation but as the rate of the HSPA bearers is not reduced any more, now the LTE connections are going to starve [7]. Without CC, the Node B defines the resource grants allocated to the bearers so that the air interface resources are not wasted, which might even increase the transport congestion.

As explained above, HSPA CC is needed but the existing solutions are causing serious fairness problems in multi-RAN systems. This paper proposes an alternative solution that achieves fair operation by adapting the rate of both HSPA and EPS bearers sharing the congested link.

III. The Centralized Congestion Control

The proposed centralized CC and resource management solution is an efficient, flexible and versatile mechanism that is capable of resolving DL and UL congestion in multi-RAN (HSPA and LTE) and homogeneous (HSPA- or LTE-only) systems, being a feasible alternative of the existing HSPA CC mechanisms. It provides the enforcement of the HSPA/LTE QoS architectures (or any other bandwidth sharing or QoS differentiation policy) and it is able to guarantee the intra- and inter-system fairness.

The architecture of the solution is shown in Fig. 2. For the sake of simplicity, the description assumes that congestion can occur only on the last mile and aggregation links, i.e., it can affect only the traffic on the S1 and Iub interfaces. This is a reasonable assumption as the backbone network is not capacity limited due to the built in redundancy. In multi-RAN or HSPA-only systems, the HSPA CC mechanisms are replaced by the centralized CC, i.e., it takes over the bandwidth control functionalities, whereas the HSPA flow control mechanisms are only responsible to grant resources to the HSPA RABs according to the needs of the packet scheduler. The solution consists of the following elements: DL congestion detection entities located in the Node Bs and eNBs; UL congestion detection entities located in the RNC and in the SAE-GW; the centralized CCE, the topology database and DL per HSPA and EPS bearer shapers located at the SAE-GW; UL per HSPA and EPS bearer shapers located at the Node Bs and eNBs, respectively. One possible mechanism to detect congestion is to use the features of the ECN (Explicit Congestion Notification) [14] but the centralized CC is expected to work with any other congestion detection method as well. Congestion is detected when the ratio of the received CE (Congestion Experienced) marked IP packets exceeds a predefined detection threshold. The benefit of the ECN is that it is an already existing standardized functionality that provides explicit congestion indication by setting the relevant fields within the IP packet header [14]. The DL congestion detection entities residing in the Node Bs/eNBs communicate directly with the CCE via CI messages. The CCE identifies the source of the CI messages indicating DL congestion based on the ID of the sender coded into the message. For detecting UL congestion at the Iub interfaces, the CCE uses the services of the detection entity residing at the RNC, which sends a separate CI message per each Iub interface whenever it detects congestion. The ID of the Node B with congested Iub interface is encoded to this message. Finally, UL congestion on the S1 interfaces is detected by the detection entity located at the SAE-GW that sends CI to the CCE.

The CCE uses a time window based congestion control algorithm. During the window, the CIs are collected and the throughput of the active bearers are measured in both directions. At the end of each time window, provided that no CI was received, the CCE starts a new window. If a new CI was received, the CCE performs a CC action, consisting of the following four procedures: (a) congested link identification; (b) bandwidth recalculation for those interfaces that share the congested link; (c) sending the Bandwidth Allocation (BA) commands to the corresponding per bearer shapers; (d) execution of the BA commands. One CC action handles one congested link; if more congested links are identified by the CCE, a separate CC action is performed for each identified congested link. In this paper, the time interval in which the CC actions are performed is referred to as a CC period. During the CC period, no new CIs are accepted from the same source, i.e., the received CIs are ignored by the CCE.

Congested link identification. The CCE uses a topology database, which contains two entries for each link in the

 ISBN: 978-1-61208-203-5

network topology, one entry for each direction. Each entry contains the link ID, denoted by k, the link capacity C_k and a list of Node Bs/eNBs whose Iub or S1 traffic is routed via link k in the corresponding direction. For the sake of simplicity, it is assumed that each Node B/eNB has one S1 or Iub interface and that the links are symmetric, i.e., the link capacity is the same in both directions. The topology database is continuously updated by the CCE, i.e., entries are added or removed as the routes of the S1 and Iub change at the end of each window. To identify the congested link(s), the CCE ranks the links based on their likelihood of being congested. For that, the CCE uses a heuristic scoring method by which the following principles are considered: (a) link k is considered to be congested if CI has arrived from a Node B/eNB served by link k and $l_k > l^{(TH)}$, where $l_k = \tau_k/C_k$ is the load of link k, τ_k is the aggregated throughput of the active bearers routed through link k and $l^{(TH)}$ is a predefined threshold for the link load; (b) if for a given CI multiple links meet these conditions, the link at higher aggregation level is considered to be the congestion point, which provides a faster convergence to a congestion free state and better inter-node fairness. The aggregation level is represented by the number of served Node Bs/eNBs, denoted by $n^{(N)}$. If each CI received during the window resulted in the selection of a separate link, it does not matter which link is selected first because the others will also be selected later in the same CC period. The CCE calculates the score s_k of each link according to (1) and selects link k with the highest s_k, i.e., considers that link as being congested.

$$s_k = \begin{cases} n^{(N)} & \text{if } l_k > l^{(TH)} \text{ and CI is received} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Resource recalculation. After link k is selected, the CCE recalculates the shaping rates of the corresponding active bearers by considering the available resources, their QoS parameters and the fairness policies:

$$R_i = r \cdot \frac{w_i}{\sum_{j=1}^{n_k^{(b)}} w_j} \cdot C_k \quad \text{where} \quad r < l^{(TH)} < 1 \quad (2)$$

where R_i is the calculated shaping rate of bearer i, r is a multiplicative decrease factor, w_j is the weight of the bearer defined in Section II-B and $n_k^{(b)}$ denotes the number of bearers in the Node B/eNB set served by link k.

Sending the BA command. Based on the R_i shaping rates of the bearers calculated in the previous step, the bandwidth allocated to each affected Node B/eNB can be determined by summing up the rate of the bearers being served by the corresponding Node B/eNB. The bandwidth allocated to a Node B/eNB must not exceed the minimum of the link capacities along the route from the GW to the corresponding Node B/eNB. If this condition is not met, the minimum of link capacities must be allocated as the bandwidth to the Node B/eNB and the deficit must be reshared among the other Node Bs/eNBs. Here this method is referred to as deficit resharing. The allocated bandwidth is sent via the BA commands to the per bearer shapers.

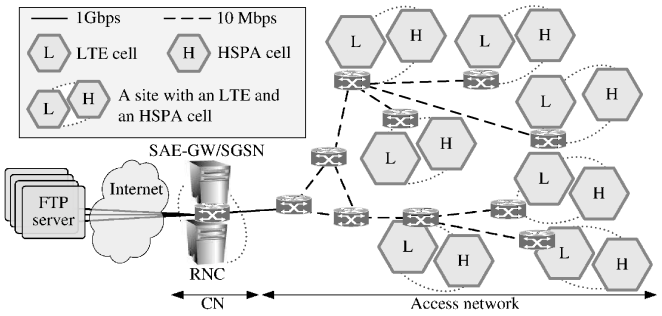


Figure 3. Simulation topology

Execution of the BA command. The shapers distribute the allocations among the active bearers (using a formula analogous to (2)) and initiate a prohibit timer. If the timer expires and no BA is received, the shapers start to increase the rate of the active bearers with an additive increase mechanism, clocked by the prohibit timer.

After BA commands are sent, the CIs of the Node Bs/eNBs served by the congested link are ignored. If there are remaining links with unhandled congestion, the CCE continues with new CC actions until all the congested links are handled, which is indicated by all link scores being zero. At that time, the CCE starts a new time window, accepting CIs again.

It is ensured by (1) that links with low load, which are not congested, are never selected by the CCE. It is also ensured that if the GW receives a CI, the CCE will perform a CC action, which will resolve the congestion by reshaping the corresponding bearers within a few CC periods. In addition, the deficit resharing mechanism ensures that the CC action does not induce further congestion on other links.

IV. Performance Evaluation

The performance of the centralized CC algorithm was analyzed with simulations. The simulation model implements in detail the UP protocols and interfaces (shown in Fig. 1), the Radio Layer 2 (PDCP/RLC/MAC) protocols, the transport network layer protocols (Iub: UDP/IP/Ethernet, S1, X2 and Iu-PS: GTP/UDP/IP/Ethernet, etc.) and the mobility procedures including the relevant control messages. Intra-system HOs (handovers) are modeled: hard HOs (HDSPA and LTE) and soft HOs (HSUPA). The details of the simulation models and the radio interface model can be found in [3].

The simulated logical topology (Fig. 3) consists of seven multi-RAN sites, each deployed both with an LTE eNB and a Node B. Each eNB and Node B is simulated with a one cell one sector configuration. The HSPA users are connected via HS-DSCH in DL and via E-DCH in UL to the RNC, whereas the LTE users are connected via DL-SCH (Downlink Shared Channel) in DL and via UL-SCH (Uplink Shared Channel)

 ISBN: 978-1-61208-203-5

in UL to the LTE eNBs. The SGSN, the SAE-GW, the MME and the RNC are considered to be co-sited. The FTP servers are connected to the SAE-GW/SGSN via the Internet. The CN consists of the RNC and the SAE-GW/SGSN/MME, interconnected through the core router. The access part of the network has a tree topology with 10 Mbit/s links. The access network is connected to the CN with a 1 Gbit/s link. The link capacities were selected in such a way that only the access links can be congested. The performance of the solution was analyzed by considering both DL (i.e., file downloads) and UL (i.e., file uploads) dominated traffic mix. Accordingly, at each simulation case, the users had either continuous file downloads or uploads to/from the FTP servers (located at the Internet) depending on the traffic mix. The TCP stack implemented the New Reno variant with 64 kB maximum advertised window size. At the transport layer, each bearer was mapped to the same PHB (Per-Hop Behavior). The minimum/maximum thresholds and the maximum drop probability parameters of the RED algorithm were set to 0.5, 1.0 and 0.1, respectively. At simulation start, the users were distributed evenly among the cells. In order to evaluate the performance of the solution under low, moderate and high load, the amount of users per cell was increased from 2 up to 6 in step of 1 that resulted in five distinct cases. The total amount of active HSPA and LTE users was equal in each case. The mobility model was random waypoint with velocity of 3 km/h. Users were executing intra-system HOs triggered according to the mobility procedures; therefore, the amount of users connected to a given Node B/eNB was changing depending on their actual location.

Three system alternatives were evaluated: (a) with no CC at all except the end-to-end TCP CC; (b) with HSPA CC only and (c) with centralized CC. When there is no CC in the system, HSPA users (both in DL and UL) receive much better service; their average throughput is at least 2.5 times of the throughput of the LTE users (Fig. 4). The reason is that the rate of FTP connections over LTE is reduced by the TCP CC whenever packet drops due to congestion are detected. In contrast, the RLC AM entity retransmits the dropped packets of the FTP connections over HSPA, which prevents TCP CC actions. The transport links are dominated by the HSPA connections that can achieve reasonable throughput whereas the RLC AM RTX rate is above 30% (Fig. 5). When there is only HSPA CC in the system, due to the shorter feedback loop, it detects congestion before the TCP CC and the rate of the HSDPA connections is reduced until their starvation (Fig. 6). This helps the LTE connections dominate the transport links. Note that in most of the cases, the HSUPA connections have lower throughput than the UL LTE connections but they are not starving. This is because the air interface capacity is narrower in UL than in DL, therefore the HSUPA and LTE air interface schedulers keep the rates of the UL flows at a lower level. Accordingly, the transport is less congested in UL than in DL.

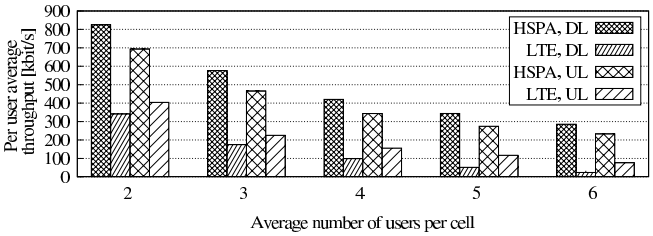


Figure 4. Per user average throughput if there is no CC in the system

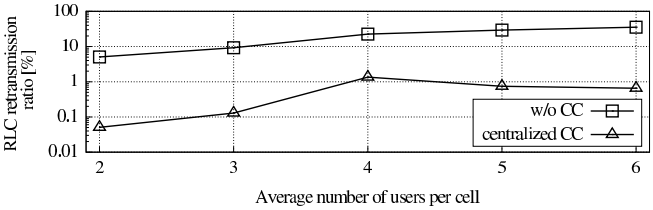


Figure 5. RLC RTX ratio over the Iub interface in DL. Results with only HSPA CC are omitted as HSPA connections are starving in that case.

The proposed centralized CC mechanism provides good level of service for both HSPA and LTE connections; their DL and UL average throughput is approximately the same (Fig. 7). The RLC AM RTX ratio is kept at reasonably low level (Fig. 5). If there is no CC or only HSPA CC used in the system, the intra-system fairness (evaluated with Jain's fairness index [15]) is poor in DL and a bit better in UL whereas the centralized CC is able to guarantee fair system operation both in DL and UL (Fig. 8).

The capability of harmonized QoS enforcement of the centralized CC was investigated in a scenario with two common QoS classes: high priority (HP) and low priority (LP). The SPI/QCI weights of the HSPA/LTE connections (bearers) were set to $wSPI_{HP} = wQCI_{HP} = w_{HP} = 2$ and to $wSPI_{LP} = wQCI_{LP} = w_{LP} = 1$. The meaning of the weights is defined in Section II-B. Three simulation cases were considered with 2 (1 HP, 1 LP), 4 (2 HP, 2 LP) and 6 (3 HP, 3 LP) users per cell according to low, moderate and high load (as before, the amount of HSPA and LTE users was equal). The results show that the centralized CC algorithm is able to provide harmonized QoS enforcement in case of DL traffic (Fig. 9): $\tau_{HP}/w_{HP} \approx \tau_{LP}/w_{LP}$ both in case of HSPA and LTE under each load (low, moderate and high), which is according to the expectations defined in Section II-B. Due to space limitations, the UL results, which are similar to DL ones, are not included.

V. Conclusion

This paper provides an overview of the aspects of QoS and fairness enforcement in multi-RAN systems sharing a common packet based transport network. At congestion, the users experience a fairness problem caused by technological and architectural differences of WCDMA/HSPA and LTE systems. WCDMA/HSPA networks with Radio

 ISBN: 978-1-61208-203-5

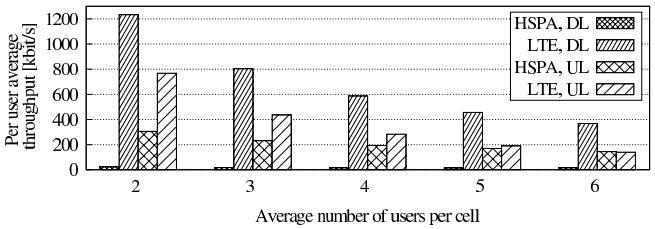


Figure 6. Per user average throughput with HSPA CC

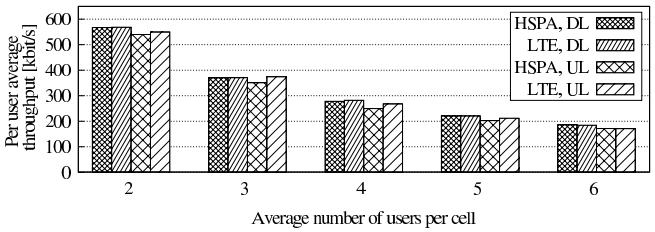


Figure 7. Per user average throughput if the centralized CC is used

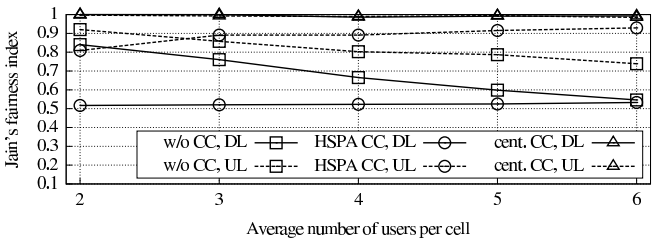


Figure 8. Jain's fairness index in DL and in UL. Index value close to 1 indicates high level of fairness.

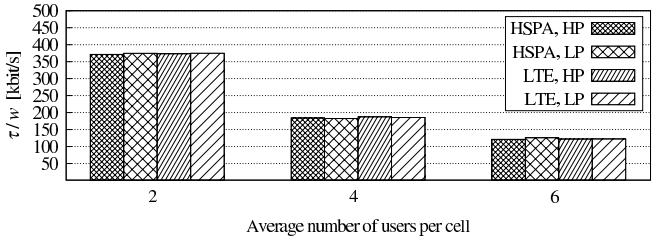


Figure 9. QoS differentiation capability of the centralized CC

Network Layer protocols such as RLC terminated at the RNC require special CC mechanisms in order to avoid performance degradation due to RLC AM RTXs over the Iub triggered by packet discards at transport congestion. The existing solutions work well in homogeneous HSPA systems but due to their intrinsic properties, they fail in multi-RAN environments. The centralized CC proposed by this paper provides a viable solution to the fairness problem combined with an efficient congestion handling and harmonized QoS differentiation capability, regardless of the traffic type. The solution is feasible both for DL and UL congestion control and can be applied in homogeneous HSPA or LTE networks as well. Simulation results confirmed that with centralized CC, the available bandwidth is shared in a fair way among the HSPA/LTE bearers regardless of the level of congestion. High fairness index, low RLC AM RTX rate and almost ideal QoS differentiation prove the superiority of the solution.

REFERENCES

[1] 3GPP, "Iub/Iur congestion control," TR 25.902 V7.1.0, 2007.

[2] ——, "Radio Link Control (RLC) protocol specification," TS 25.322 V10.1.0, 2011.

[3] L. Kőrössy and Cs. Vulkán, "QoS Aware HSDPA Congestion Control Algorithm," in *Proc. of IEEE International Conference on Wireless and Mobile Computing*, Avignon, France, Oct. 2008, pp. 404–409.

[4] B. Suter *et al.*, "Design considerations for supporting TCP with per-flow queueing," in *Proc. of INFOCOMM'98*, vol. 1, San Francisco, USA, Apr. 1998, pp. 299–306.

[5] "Optimizing global mobility through seamless coexistence and evolution of GSM, WCDMA and LTE," White Paper, Ericsson, Tech. Rep., Feb. 2009.

[6] K. D. Singh and D. Ros, "TCP-Friendly Rate Control over High-Speed Downlink Packet Access," in *Proc. of 12th IEEE Symposium on Computers and Communications*, Aveiro, Portugal, Jul. 2007, pp. 515–521.

[7] Cs. Vulkán and B. Héder, "Congestion Control in Evolved HSPA Systems," in *CD Proc. of IEEE 73rd Vehicular Technology Conference (VTC'11 Spring)*, Budapest, Hungary, May 2011, Paper No.: 1081710.

[8] 3GPP, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," TS 24.301 V11.0.0, 2011.

[9] K. I. Pedersen *et al.*, "Overview of QoS Options for HSDPA," *IEEE Commun. Mag.*, vol. 44, no. 7, pp. 100–105, 2006.

[10] H. Holma and A. Toskala, Eds., *LTE for UMTS: OFDMA and SC-FDMA Based Radio Access*. John Wiley & Sons, 2009.

[11] T. E. Kolding, "QoS-Aware Proportional Fair Packet Scheduling with Required Activity Detection," in *Proc. of IEEE 64th Vehicular Technology Conference (VTC'06 Fall)*, Montréal, Canada, Sep. 2006, pp. 1–5.

[12] S. Nádas *et al.*, "Providing Congestion Control in the Iub Transport Network for HSDPA," in *Proc. of GLOBECOMM'07*, Washington D.C., USA, Nov. 2007, pp. 5293–5297.

[13] 3GPP, "UTRAN Iub/Iur interface user plane protocol for DCH data streams," TS 25.247 V11.0.0, 2011.

[14] K. Ramakrishnan *et al.*, "The Addition of Explicit Congestion Notification (ECN) to IP," *IETF RFC 3168*, Sep. 2001.

[15] R. K. Jain *et al.*, "A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems," DEC-TR-301, Digital Equipment Corporation, Tech. Rep., September 1984.

 ISBN: 978-1-61208-203-5

A Fast and Efficient Key Agreement Scheme for Wireless Sensor Networks

Mee Loong Yang*, Adnan al-Anbuky†, and William Liu‡
*School of Computing & Mathematical Sciences
Email: bobby.yang@aut.ac.nz
†School of Engineering
Email: aalanbuk@aut.ac.nz
‡School of Computing & Mathematical Sciences
Email: william.liu@aut.ac.nz
Auckland University of Technology, New Zealand

***Abstract*—The Blom's scheme for key agreement between pairs of nodes requires exchange of a small amount of bits, uses simple computations, and also authenticates each other. This makes it attractive for use in Wireless Sensor Networks but, in its original form, it has limitations because of the contending requirements for large pairwise keys and limited memory in the nodes. Our implementation of the Blom's scheme uses multiple keys, enabling it to derive large pairwise keys using the limited memory resources, while retaining all the desirable features of speed, compactness, and low energy usage. We implemented our scheme in a MICAz mote and present some experimental results on the memory, computation time, and energy requirements. We compared the performance with other public key cryptographic methods used in WSN. Our scheme, using 382 bytes of RAM, was able to compute 128-bits pairwise keys in times ranging from 34 ms to 1.9 s for networks with capture thresholds of 32 and about 2,000 nodes respectively.**

Keywords- Blom's scheme; ad hoc networks; security; wireless sensor networks; key pre-distribution.

I. INTRODUCTION

A wireless sensor network deployed in open environments can be easily attacked. The radio communications can be eavesdropped, and an adversary can even inject malicious packets into the network. One important part in the chain of defence is to protect the communication channel between the nodes. This means encrypting the messages to protect confidentiality. In addition, the receiving node must be able to verify that the received messages are intact, fresh, and really originates from the claimed sender. Proven cryptographic techniques can be used to achieve these requirements. These techniques rely on the use of secret keys known only to the communicating parties. These keys must be distributed to the nodes in a secure manner.

One approach to key distribution requires the base station to generate all the keys which the nodes will use. Many early works take this approach using a global master key, such as in SPINS [1], and LEAP [2]. Obviously, if the master key is compromised, so is the whole network.

A more secure approach would use unique pairwise keys between nodes to limit the impact of any key compromise. However, in ad hoc networks, the pairwise relationships cannot be predetermined. In a network with N nodes, each node has $(N-1)$ pairwise keys with all its neighbours and a large amount of memory would be required to store them all. Even if this is possible, a single node compromised also affects the whole network. When deployed, nodes only need to have common keys with their immediate neighbours. Therefore, a probabilistic approach can be used. This was done in [3] where nodes are given subsets of keys from the global key pool. After deployment, pairs of nodes discover shared keys to establish secure links. Even without shared keys, pairs of nodes can attempt to use secured mutual intermediary nodes to establish a secure link. If one node becomes compromised, it only impacts on part of the network.

A different approach requires pairs of nodes to contact a trusted centre to provide their pairwise keys. Each node needs only to be provided with a pre-shared key with the trusted centre. This is the Key Distribution Centre (KDC) scheme used widely in protocols such as Kerberos [4]. In ad hoc mobile networks, this scheme is of limited use due to the requirement for the trusted centre to be reachable at all times.

Key pre-distribution schemes, on the other hand, pre-deploy nodes with keying material which they will use to derive pairwise keys with their neighbours. Such schemes commonly use public key cryptography (PKC). An example is the Diffie-Hellman (DH) protocol used widely in wired networks. For WSN, the Elliptic Curve DH (ECDH) is promising due to its less demand on resources compared to other PKC methods. Several implementations have been studied as in [5], [6], [7], [8]. These methods enable nodes to derive a common secret key by exchanging some information over the insecure channel. There is also the need for nodes to authenticate each other, usually using a certificate such as in the ECDH-ECDSA protocol implemented in [9], or using the Menezes-Qu-Vanstone (ECMQV) protocol [10], and that based on the ElGamal scheme [11]. We shall refer to these ECC methods later when comparing their performances with our scheme.

An interesting key pre-distribution scheme, which has

 ISBN: 978-1-61208-203-5

implicit authentication, was proposed by Rolf Blom [12]. This scheme enables pairs of nodes to compute a common secret pairwise key after exchanging a small number of bits. The computation uses simple arithmetic operations and requires only a few steps. This makes it attractive for use in WSN. However, as shown later, in its basic form, it has limitations due to the contending requirements for small storage, large pairwise key sizes, and large number of nodes in the network. This paper describes our modifications to the scheme enabling it to derive large pairwise keys after exchanging a small number of bits. We also report the results of our implementation in a MICAz mote.

This paper is structured as follows. Section II presents the background and some related works. Section III describes our modifications to the Blom's scheme using multiple-keys. Next, in Section IV, we present the experimental results of our implementation in a MICAz mote. We discussed some of the security features and applications of our scheme in Section V. Then, we made some comparisons with other PKC methods in Section VI, and finally we gave our conclusion in Section VI.

II. Related work

A. Background: Blom's scheme

In this scheme, the base station generates a secret $(m \times m)$ symmetric matrix $\mathbf{S}$. Each node is assigned a unique $(m \times 1)$ column vector e.g., $\mathbf{V}_A$, and $\mathbf{V}_B$, for nodes A and B, respectively. These vectors are called the node's *public identifiers (IDs)* or *public vectors*. The base station then computes and stores in each node their private keys which are row vectors $K_x = \mathbf{V}'_x \cdot \mathbf{S}$. . These public IDs and private keys form the keying material for the node.

Any two nodes can derive a pairwise secret key between them. For example, between nodes *A* and *B*, the nodes exchange their public IDs and then compute a common key, K_{AB}.

$$\begin{aligned} \text{Node A: } K_{AB} &= (\mathbf{V}'_A \cdot \mathbf{S}) \cdot \mathbf{V}_B \\ \text{Node B: } K_{BA} &= (\mathbf{V}'_B \cdot \mathbf{S}) \cdot \mathbf{V}_A \\ K'_{BA} &= (\mathbf{V}'_B \cdot \mathbf{S} \cdot \mathbf{V}_A)' = \mathbf{V}'_A \cdot \mathbf{S}' \cdot \mathbf{V}_B \end{aligned}$$

Since $\mathbf{S}$ is symmetric, $K_{AB} = K'_{BA}$. An important feature is that success in deriving a common key authenticates the nodes to each other since this requires valid private keys provided by the base station.

In the key agreement process, the public IDs can be transmitted in clear text. The private keys must be kept secret. If an adversary captures the nodes, they may be able to obtain their keying material. If sufficient number of nodes are captured, these information can be used to derive the secret matrix $\mathbf{S}$ and hence break the whole scheme. For this to be possible, the number of nodes captured must be $\geq m$, and they must all have unique linearly independent public vectors. The system is said to be $(m-1)$ secure, i.e., a coalition of $(m-1)$ or less nodes cannot pool their keying material to derive the pairwise key of any other pair of nodes [13].

The column vectors of the Vandermonde matrix $\mathbf{V}$ given below is a suitable choice for public IDs since all s_k are distinct. The base station assigns to each node, one of the columns of $\mathbf{V}$. In practice, every node only need to be bootstrapped with the seed s_k from which to generate the public ID vector.

$$\mathbf{V} = \begin{bmatrix} 1 & \cdots & 1 & \cdots & 1 \\ s_1 & \cdots & s_k & \cdots & s_N \\ {s_1}^2 & \cdots & {s_k}^2 & \cdots & {s_N}^2 \\ \vdots & \ddots & \vdots & \ddots & \cdots \\ {s_1}^{m-1} & \cdots & {s_k}^{m-1} & \cdots & {s_N}^{m-1} \end{bmatrix}$$

Security parameters: If the network is $(m-1)-secure$, i.e., the matrix size is $(m \times m)$, and the size of each element of $\mathbf{S}$ is b -bits, then for a fully secure system,

$$\begin{aligned} \text{Max network size } Q_N &= (m-1) \text{ nodes} \\ \text{Public-ID seed size, } Q_v &= b \text{ -bits} \\ \text{Private-Key size, } Q_{ku} &= m \times b \text{ -bits} \\ \text{Pairwise key size, } Q_{kpair} &= b \text{ -bits} \end{aligned}$$

The pairwise key should be large, 64 -bits or larger, leading to large b. While the MICAz motes are able to work with 8, 16, 32 and 64 -bit data sizes, larger sizes for the vector elements would make the computations more complicated.

B. Related Works

The number of nodes in a fully secure network can be increased by using multiple key spaces. In [14], ω key spaces are generated and each node is given a sub-set of τ randomly chosen keys from ω. After deployment, nodes discover their common keys and use the Blom's scheme to form pairwise keys. The scheme uses a similar idea to the probabilistic scheme of Eschenaeur-Gligor [3] where nodes are given a random set of keys from a global key space. In these schemes the aim is to achieve full connectivity, but not necessarily complete connectivity like a full mesh. Another approach also uses Blom's scheme with multiple key spaces to improve resistance to the Sybil attack [15].

In [16], the scheme for a clustered topology is proposed. Here, the cluster-heads implement the Blom's scheme to derive pairwise keys among themselves. Non cluster-head nodes do not implement the Blom's scheme. Instead, they store a pre-computed secret key K_i for use with a cluster head. Prior to deployment, the base station computes the pairwise keys of this node with a certain number of associated cluster-heads. These are then combined into a secret key K_i and stored in the node, together with the identities (IDs) of the associated cluster-heads. When a node needs to establish a secure link with a physical cluster-head,

 ISBN: 978-1-61208-203-5

it sends its own ID and the IDs of its associated cluster-heads. The physical cluster-head forwards the node's ID to the associated cluster-heads to compute the pairwise keys using Blom's scheme and thereby derives the secret key K_i. In this way, non-cluster head nodes store minimum keying material and do not need to perform any key computation computation. Instead, these are delegated to the cluster heads which carry the additional load of communicating with other cluster heads to derive the key with a non cluster-head node. The network size would still be limited to the $(m-1)$ nodes for a fully secure network. Since cluster-heads establish pairwise keys among themselves using the basic Blom's scheme, the key size and memory requirements, and network size would still be limited to the original scheme.

C. Our Main Contributions

We modified the Blom's scheme using multiple keys such that it is able to derive large pairwise keys of up to hundreds of bits using 16-bit data sizes. It requires very little RAM for the computation while retaining all the benefits of the basic scheme i.e., mutual authentication, few exchanged bits, simple computations, fast, and consumes little energy. This makes it especially suitable for dynamic B sensor networks where the nodes are highly mobile and key computation and authentication must be achieved quickly and cheaply. It is also scalable for fixed cluster based topologies.

III. Blom's scheme with Multiple-keys

In our modification to the scheme, the base station generates N secret symmetric $(m \times m)$ matrices $S_1, S_2, \cdots, S_N$, and one Vandermonde matrix $\mathbf{V}$, over the prime field $GF(p)$, where p is the largest prime $< b$ -bits. The number of bits, b is thus the data size used in the system.

Public IDs

Each node is given one unique set of vectors comprising N vectors of $\mathbf{V}$, called its *public ID vectors*. Since the elements of each vector is given by s_{ki} for $i = 0, \cdots, m-1$, an ID vector can be generated by anyone given s_k. Each node's public ID vectors can be simplified to the set of seeds $\{s_{k1}, s_{k2}, \cdots, s_{kN}\}$. We call this set the node's public *ID-tag*.

Private keys

For each node, the base station computes the set of private vectors or *private keys*, using all permutations of the secret matrices S_i and its public ID vectors. For node x, its private key set consist of N^2 separate $(1 \times m)$ row vectors given by,

$$K_{uxij} = \mathbf{V}'_{xi} \cdot S_j \quad \text{for } i, j = 1, \cdots, N$$

Pairwise Key Derivation

Consider two nodes *A* and *B* attempting to form pairwise key. Each has in their possessions, their public Id-tags and private keys:

$$\begin{aligned} \text{Node A:} \quad & \{s_{A_i}\}, \{K_{uAj}\} \\ \text{Node B:} \quad & \{s_{B_i}\}, \{K_{uBj}\} \\ \text{for} \quad i = 1, \cdots, N \text{ and } & j = 1, \cdots, N^2 \end{aligned}$$

To derive their pairwise key they exchange their public ID-tags.

$$\begin{aligned} A \rightarrow B: \quad & \{s_A\} \\ A \leftarrow B: \quad & \{s_B\} \end{aligned}$$

Each node after receiving the other node's public ID-tag would generate that node's ID vectors, e.g., node B would be able to generate A's public vectors $V_{A_1}, V_{A_2}, \cdots, V_{A_N}$

Using all permutations of the public ID vectors and its own private keys, each node computes a set of secret numbers:

$$\begin{aligned} \text{Node A:} \quad & K_{uAj} \cdot \mathbf{V}_{B_i} \\ \text{Node B:} \quad & K_{uBj} \cdot \mathbf{V}_{A_i} \\ \text{for} \quad i = 1, \cdots, N \text{ and } & j = 1, \cdots, N^2 \end{aligned}$$

Both parties would obtain a set of identical N^3 secret numbers, not necessarily arranged in the same order. Each number is of size b -bits. Using an agreed rule, sufficient numbers are chosen and concatenated together to form the pairwise key K_{ABpair} of the desired size. For example, a simple rule would be for each node to sort the numbers in descending order and concatenate the first 8 numbers to form a pairwise key of size $8 \times b$ -bits.

The key computation code is very simple and compact as shown in the pseudo code in Listing 1.

Listing 1. Pseudo code for pairwise key computation

```
generateKeyPair(){
  for (each public_ID-tag value) {
    generate public_vector;
    for (each private_key) {
      multiply with public vector;
      save result in SecretNumbers;
    }
  }
  sort SecretNumbers;
  select numbers from SecretNumbers;
  concatenate to form pairwise key;
}
```

 ISBN: 978-1-61208-203-5

Some parameters

Pairwise key size: The maximum pairwise key size, if all the secret numbers are used, is bN^3 -bits. With $N = 2$ and using 16 -bit data sizes, we have 128 -bits key sizes which is more than enough if it is used in a secure algorithm such AES.

Memory requirement: The private key requires the largest amount of memory. This is static data and can be assigned to program memory in the MICAz mote. The amount of RAM required include those for some counters, the pairwise key, some temporary data, and the N^3 secret numbers. In total, the RAM requirement is very small indeed as shown later.

Computation time: The main part of the computation involves modulo multiplication and addition of the m elements of the N public ID vectors and N^2 private key vectors. In total, there are mN^3 such operations.

IV. Experimental Results

We implemented the scheme in a MICAz [17] mote using TinyOS [18] for the case using 16 -bits data size. We also estimated the energy consumed for the key computation process. In our experiment, the code was kept to the bare minimum for key computation and turning on the LEDs at the start and end of process. Even though the modulo operations took a significant amount of time, no attempt was made to optimise it, and only the standard libraries were used for all operations.

The private keys were installed in the program code. The public ID-tag exchange was not implemented and was merely simulated by storing the public ID-tag of the simulated neighbour as a variable in the node.

When the program runs, it lights up an LED, computes the pairwise key, and lights up another LED on completion. The time taken for key computation was estimated by timing the 100 iterations of the computation. The power supply to the node was regulated at 3.1 V and the average current during computation was measured to be 8.7 mA.

Performance and analysis

The results for memory requirements, key computation time, and estimated energy consumed are shown in Table I.

Memory requirements: The MICAz mote has 4 kB RAM and 128 kB flash memory for program. In our implementation, we stored the private keys in program memory leaving more RAM for the variables. The ROM and RAM requirements were outputs from the TinyOS-2.1.1 compiler.

The private key vectors has mN^2 elements, each of 16 -bits. From the results, the ROM storage requirements in bytes was, as expected, a linear relationship as given below.

$$Q_O = 2.012\ mN^2 + 7010 \tag{1}$$

The number of variables in RAM was fixed except for the N^3 secret numbers. From the experimental results, the following relationship was obtained for the RAM storage requirements, in bytes:

$$Q_R = 8\ N^3 + 318 \tag{2}$$

The key computation process involves multiplying the m elements of N public ID vectors with the m elements of N^2 private key vectors, plus one sorting and selection operation. It was a linear relationship between the computation time and mN^3. From the results we obtained the following relationship for key computation time in ms,

$$t = 0.0514\ mN^3 + 24.60 \tag{3}$$

Computation energy: The average current drawn during the computation from the 3.1 V regulated power supply was measured to be about 8.7 mA. Using this, the estimated energy in mJ, used for computation was estimated as $3.1 \times 0.0087 \times t$, also shown in Table I.

Design example: The above results can be used for design purposes. For example, we have a network of 500 nodes and we wish to select the parameters and estimate the computation time and memory requirements. For a fully secure network, the number of nodes is $< \frac{m}{N}$. Trying with $N = 2$, a suitable choice is $m \geq 2 \times 500 = 1000$. Using a slightly larger value, $m = 1,024$, and $N = 2$ from (3) the key computation time would be 446 ms. The storage requirements, from (1) would be 15,252 bytes ROM, and from (2) gives 382 bytes RAM.

The results of an actual implementation in a MICAz mote was 15,282 bytes ROM, 384 bytes RAM, and the computation time was 479 ms.

V. Security features and applications

Brute force attacks: To attempt to guess the pairwise key or the private keys would be infeasible as these are large, at least 64 -bits, and hundreds of bits respectively.

Node capture

Nodes can be physically captured and sensitive information extracted unless tamper proof hardware mechanisms can be incorporated. This would increase the cost and probably not be viable except for critical situations. If such mechanisms are available, the scheme would be very attractive for highly mobile, ad hoc networks. For example, using a small $m = 24$ and $N = 2$ in (3), pairs of nodes can derive keys in 34.5 ms requiring about 0.93 mJ of energy.

Currently, motes do not have adequate tamper protection and an attacker with the appropriate skills and resources can easily obtain the memory contents from motes like the MICA2 [19], and TelosB [20].

The $(m - 1)$ -secure property of the Blom's scheme still applies in our multiple-key case. If an attacker manages to obtain m sets of linearly independent public IDs and the associated private keys, it is possible to craft valid public

 ISBN: 978-1-61208-203-5

	Number of keys, N			
	1	2	3	4
matrix size: 64 × 64				
ROM (bytes)	6,888	7,678	8,312	9,206
RAM (bytes)	326	382	534	830
time (ms)	9	34	97	231
Energy (mJ)	0.24	0.92	2.62	16.23
matrix size: 128 × 128				
ROM (bytes)	7,016	8,192	9,466	11,260
RAM (bytes)	326	382	534	830
time (ms)	15	63	176	399
Energy (mJ)	0.40	1.70	4.75	10.76
matrix size: 256 × 256				
ROM (bytes)	7,274	9,210	11,772	15,358
RAM (bytes)	326	382	534	830
time (ms)	30	124	332	734
Energy (mJ)	0.81	3.34	8.95	19.8
matrix size: 1,024 × 1,024				
ROM (bytes)	8,812	15,354	25,596	*
RAM (bytes)	326	382	534	*
time (ms)	121	480	1,275	*
Energy (mJ)	3.26	12.95	34.39	
matrix size: 4,095 × 4,095				
ROM (bytes)	14,954	39,922	*	*
RAM (bytes)	326	382	*	*
time (ms)	486	1,906	*	*
Energy (mJ)	13.11	51.40	*	*

Table I
MEMORY REQUIREMENTS, COMPUTATION TIMES, AND ESTIMATED ENERGY FOR KEY COMPUTATION USING 16-BITS DATA SIZE. * THESE EXCEEDS THE ARRAY LIMIT

IDs and private keys for any other nodes. It would also be possible to derive the secret matrices and completely break the system. In our scheme, the number of captured nodes which can compromise the network, called the "capture threshold", is $Q_c = \frac{m}{N}$.

Depending on the application, we can implement the scheme with suitable levels of security for the following topologies,

- fully secure, ad hoc, mobile, full mesh topology
- fully secure, fixed, cluster topology,
- partially secure, ad hoc, mobile, full mesh topology

1. Fully secure, ad hoc, mobile, fully mesh topology: The scheme can be directly applied in this situation. All nodes are mobile and are able to form pairwise keys with every neighbour within range. The number of exposed nodes in the network is kept to below the capture threshold. Further, to prevent the captured keying material from being used to craft another node, the nodes have unique, non-intersecting sets of public ID vectors. This means there can be at most $\lfloor \frac{m-1}{N} \rfloor$ nodes in the network. This limits the network size to about 2,000 nodes for the case of $N = 2$, and $m = 4,095$ with the key computation time of about 1.9 s.

Smaller networks with highly mobile nodes such as in sports or combat situations would specially benefit from this scheme. For example, with about 30 nodes, using $m = 64$ and $N = 2$, nodes can derive a pairwise key in 34 ms using about 0.92 mJ of energy.

2. Fully secure, ad hoc, fixed cluster topology: If the network is organised as clusters and the nodes are fixed in position, we can implement the multiple-key Blom's scheme among the cluster head nodes, restricting their number to be less than the capture threshold. The leaf nodes also uses the scheme but with a difference in that after deployment, their public ID-tags are deleted once they have established a pairwise link with a cluster head, or within a certain time window. Without the ID-tags, the private keys cannot be used to compromise the system. Thereafter, the leaf node do not implement the scheme. For example, using $m = 4095$, $N = 2$, we can have up to about 2,000 cluster heads. If each cluster head has 10 nodes attached to it, the network size would be 2,500 nodes.

3. Partially secure, ad hoc, fully mesh topology: In some situations it is required to protect the network against casual or opportunistic attacks but not necessarily against determined and fully resourced adversaries such as rival organisations. In this case, since considerable resources in effort and time is required to capture Q_c nodes, extract the keying material, and solve the matrices, it may be permissible to exceed the capture threshold. Other security features such as node capture detection if implemented, would also help to support this approach. The network can then be much larger than the capture threshold. For example, if capturing 500 nodes and extracting the keying material is considered to be infeasible, we can have a thousand or more nodes in the network. The scheme can be directly implemented and have all the benefits of the ad hoc, mobile, fully meshed topology.

VI. DISCUSSIONS AND COMPARISONS

Comparision with PKC methods

The aim of our key agreement scheme is to derive large pairwise keys with authentication. Similar schemes which can achieve this are the PKC methods which have been successfully used in wired networks. Their application in WSN has been studied, such as TinyECC [21], EC-ElGamal [7], ECMQV/ECDCH [8], [6], etc. We will make some brief comparisons with our scheme in some performance metrics important in sensor networks, i.e., energy, key computation time, and memory requirements.

Exchange of keying material: Key agreement schemes requires the exchange of some keying material. The amount of bits exchanged impacts on the energy used for the radio. For comparison we will exclude the overheads such MAC addresses, protocol headers, etc.

The DH scheme requires the two parties to exchange their public keys from which to derive a common secret. To authenticate each other, the public keys need to be signed by a trusted authority. For ECC schemes, the basic

 ISBN: 978-1-61208-203-5

components include a public key which is a point on the elliptic curve, its hash value, and the signature comprising two integers provided by the trusted authority. Using 160 -bits, an authenticated ECDH scheme would require exchange of at least 768 bits. Compared to our scheme requiring Nb -bits, this is larger by more than 10 times.

Key computation time and energy: In an optimised implementation of the ECDH scheme for WSN [6], the key computation took 710 ms, and used 17.04 mJ of energy in a MICAz mote running TinyOS. Another implementation using the EC-ElGamal scheme on a MICAz mote [7] reported 570 ms for key computation. These do not include signature verifications which would also require substantial amount of time and energy. For example, in [21], TinyECC was used to implement ECDH for key computation and ECDSA for signature verification in a MICAz mote. The results showed that with all optimizations enabled, the execution times were: ECDSA initialisation 3,393 ms, verification 2,436 ms, and for ECDH initialisation 1,839 ms, and key computation 2,117 ms. Hence, key computation and signature verification can take 4.5 seconds, after initialisation of about 5.2 seconds. Our key computation times depends on the choice of m and N as shown in Table I. In the largest case with $N = 2$ and $m = 4095$, the key computation took 1.9 seconds.

Memory requirements: The largest use of memory in our scheme is bmN^2 -bits for the private keys. This is static and can be stored in program memory. The code itself is very simple and compact requiring only a few hundred bytes of ROM storage. The total ROM memory for data code was less than 40 kB for the the largest values of $m = 4095$ and $N = 2$.

RAM storage was required only for variables such as the N^3 secret numbers, the pairwise key, and some counters. For the case of $N = 2$, only 382 bytes of RAM was required, as shown in Table I.

The PKC schemes require substantially more memory, especially RAM storage. In a MICAz mote implementing TinyECC [21], the memory requirements, with all optimisations enabled for ECDSA was 19,308 bytes ROM and 1,510 bytes RAM, for ECIES 20,758 bytes ROM and 1,774 bytes RAM, and for ECDH 16,018 bytes ROM and 1,774 bytes RAM. With all optimisations disabled, all the RAM sizes were only around 150 bytes but with consequently huge execution times, such as 31 seconds for ECDH key establishment! In [22] the code size for ECDSA as 56,600 bytes ROM and 1,700 bytes RAM.

VII. Conclusion

We presented our implementation of the modified Blom's scheme using multiple-keys which, while retaining the advantages of the basic scheme, improves it to make it very attractive for use in WSN. It is able achieve large pairwise key sizes, fast, and requires little energy and computational resources. We implemented our scheme in a MICAz mote and the results showed it be very advantageous compared other PKC schemes in terms of speed, energy, and RAM storage requirements. The network is fully secure if the number of compromised nodes do not exceed the capture threshold. The best choice was $N = 2$ keys, enabling pairwise key size of 128 bits requiring only 382 bytes of RAM. The ROM requirements are mainly for the node's private keys and its size depends on the capture threshold. In our case, the largest amount required was about 40 kB for a network with capture threshold of about 2,000 nodes. The key computation time increases as the capture threshold increases. This ranged from 34 ms for a capture threshold of 32 nodes, to 1.9 s for a capture threshold of 2,000 nodes using $N = 2$ keys.

References

[1] A. Perrig, R. Szewczyk, V. Wen, D. Culler, and J. D. Tygar, "Spins: Security protocols for sensor networks," *Wireless Networks*, vol. 8, pp. 521–534, 2002.

[2] S. Zhu, S. Setia, and S. Jajodia, "Leap: Efficient security mechanisms for large-scale distributed sensor networks," *Proceedings of the 10th ACM conference on Computer and communications security*, 2003.

[3] L. Eschenauer and V. D. Gligor, "A key-management scheme for distributed sensor networks," *In Proceedings of the 9th ACM Conference on Computer and Communications Security*, pp. 41–47, 2002.

[4] J. G. Steiner, C. Neuman, and J. I. Schiller, "Kerberos: An authentication service for open network systems." *In Proceedings of the Winter 1988 USENIX Conference. USENIX*, February 1988.

[5] M. Liu, W. Wei, and Z. Liu, "A secure key pre-distribution scheme for wireless sensor networks," *International Conference on Industrial Electronics and Applications, ICIEA.*, pp. 1762 –1768, May 2009.

[6] C. Lederer, R. Mader, M. Koschuch, J. Groschdl, A. Szekely, and S. Tillich, *Energy-Efficient Implementation of ECDH Key Exchange for Wireless Sensor Networks*. Springer Verlag, LNCS 5746, September 2009.

[7] O. Ugus, D. Westhoff, R. L. 0002, A. Shoufan, and S. A. Huss, "Optimized implementation of elliptic curve based additive homomorphic encryption for wireless sensor networks," *2nd Workshop on Embedded Systems Security - WESS'2007, Salzburg, Austria.*, October 2007.

[8] J. Groschdl, A. Szekely, and S. Tillich, "The energy cost of cryptographic key establishment in wireless sensor networks," *Proc. The 2nd ACM Symposium on Information, Compter and Communication Security*, 2007.

[9] G. de Meulenaer, F. Gosset, F.-X. Standaert, and O. Pereira, "On the energy cost of communications and cryptography in wireless sensor networks," *IEEE International Conference on Wireless & Mobile Computing, Networking & Communication*, pp. 580–585, 10 2008.

 ISBN: 978-1-61208-203-5

[10] L. E. Law, A. J. Menezes, M. Qu, J. A. Solinas, and S. A. Vanstone, "An efficient protocol for authenticated key agreement," *Designs, Codes and Cryptography*, vol. 28, no. 2, pp. 119–134, 2003.

[11] J. Zheng, J. Li, M. J. Lee, and M. Anshel, "A lightweight encryption and authentication scheme for wireless sensor networks," *Int. J. Security and Networks*, vol. 1, no. 3/4, pp. 138–146, 2006.

[12] R. Blom, "An optimal class of symmetric key generation systems," Linkopping University, Tech. Rep., 1984.

[13] A. Menezes, P. van Oorschot, and S. Vanstone, *Handbook of Applied Cryptography*. CRC Press, Inc., 1996.

[14] W. Du, S. Y. Han, J. Deng, and P. K. Varshney, "A pairwise key pre-distribution scheme for wireless sensor networks," *Proceedings of the conference on Computer and communications security*, October 2003.

[15] S.-J. Wang, Y.-R. Tsai, and J.-W. Chan, "A countermeasure against frequent attacks based on the blom-scheme in ad hoc sensor networks," *International Symposium on Wireless Pervasive Computing*, 2007.

[16] N. Chen, J.-b. Yao, and G.-j. Wen, "An improved matrix key pre-distribution scheme for wireless sensor networks," *International Conference on Embedded Software Systems*, p. 4045, 2008.

[17] *Memsic Corporation, MICAz Datasheet*. [Online]. Available: http://www.memsic.com/products/wireless-sensor-networks/wireless-modules.html, Retrieved: April 12, 2012

[18] P. Levis, *TinyOS programming*, 2006. [Online]. Available: http://csl.stanford.edu/ pal/pubs/tinyos-programming.pdf. Retrieved: April 12, 2012

[19] C. Hartung, J. Balasalle, and R. Han, "Node compromise in sensor networks: The need for secure systems," Department of Computer Science, University of Colorado at Boulder, Tech. Rep., January 2005.

[20] A. Becher, Z. Benenson, and M. Dornseif, "Tampering with motes: Real-world physical attacks on wireless sensor networks," *Proceedings of the Third international conference on Security in Pervasive Computing*, vol. 3934, pp. 104–118, 2006.

[21] A. Liu and P. Ning, "Tinyecc: A configurable library for elliptic curve cryptography in wireless sensor networks," *in Proceedings of the 7th International Conference on Information Processing in Sensor Networks*, pp. 245–256, April 2008.

[22] H. Wang and Q. Li, "Efficient implementation of public key cryptosystems on mote sensors (short paper," in *In International Conference on Information and Communication Security (ICICS), LNCS 4307*, 2006, pp. 519–528.

 ISBN: 978-1-61208-203-5

Interactive Remote Authentication Dial In User Service (RADIUS) Authentication Server Model

Rohan Deshmukh
CISCO Systems Inc., India
email: rodeshmu@cisco.com

***Abstract*—Normally, RADIUS servers are passive servers, i.e., they act only on requests received from Network Access Server. In a system where there are multiple servers configured in round-robin fashion, if some of the servers go down, it takes more time to reach the actual active server after retransmissions to the non-responsive server get exhausted. Here, we present a new approach to make RADIUS Server more Interactive Server. It sends ACTIVE-Request to the Network Access Server once it becomes active and DEAD-Request once it becomes non-responsive.**

Keywords-RADIUS; NAS; ID; Attributes.

I. INTRODUCTION

RADIUS servers are being used for AAA (Authentication, Authorization and Accounting) purpose [1]. With manual intervention, RADIUS server can send CoA (Change of Authorization) and DM (Disconnect Message) to the Network Access Server (NAS) [2]. The RADIUS client function may reside in a Gateway GPRS Support Node (GGSN). When the GGSN receives a Create PDP Context request message, the RADIUS client function may send the authentication information in the request "Access-Request" to an authentication server, which is identified during the Access Point Name provisioning [3]. The NAS sends an Access-Request packet to the RADIUS Server with NAS-Identifier, NAS-Port, User-Name and User-Password. The RADIUS server then sends back either an Access-Accept or Access-Reject based on whether the response matches the required value, or it can even send another Access-Challenge. Figure 1a describes this.

If the RADIUS server does not send any response, the NAS re-transmits the request to the same server without change in its attributes like Request Authenticator, ID, and source port. If any attributes have changed, a request is generated with new Request Authenticator and ID. Use of Status-Server Packets in the RADIUS protocol is mentioned in [4]. But, it is for *clients* to query the status of a RADIUS server. While the Status-Server (12) code was defined as experimental in [1], Section 3, details of the operation and potential uses of the code are not provided.

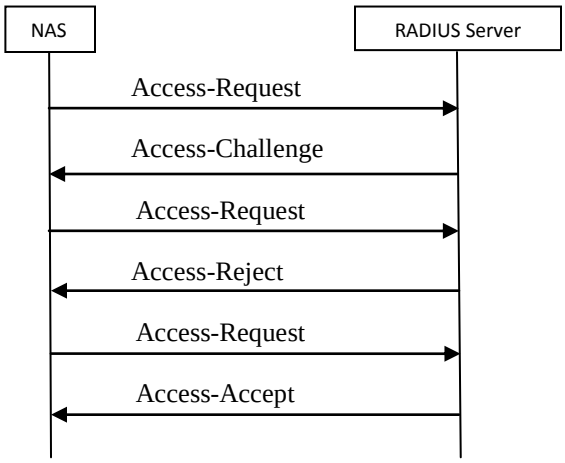


Figure 1a. RADIUS Auth Messages

The RADIUS server can send CoA and DM requests only. CoA-Request packets contain information for dynamically changing session authorizations. The NAS responds to a CoA-Request sent by a RADIUS server with a CoA-ACK if the NAS is able to successfully change the authorizations for the user session, or a CoA-NAK if the request is unsuccessful. A Disconnect-Request packet is sent by the RADIUS server in order to terminate a user session on a NAS to discard all associated session context. Figure 1b describes this.

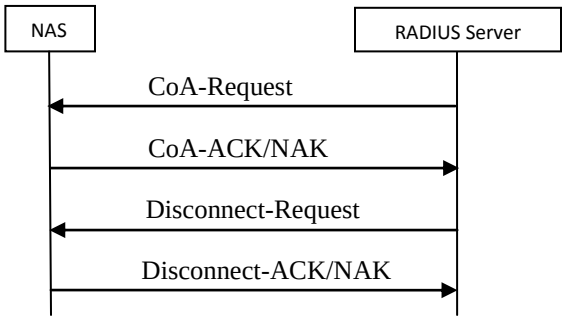


Figure 1b. RADIUS CoA/DM Messages

 ISBN: 978-1-61208-203-5

Selection of RADIUS server is based on the following algorithms:

- Round-robin: designates that the context should load-balance sending data among all of the defined RADIUS servers in a cyclic manner.
- RADIUS first-server: designates that context sends data to the RADIUS server with the highest configured priority. In the event that this server becomes unreachable, data is sent to the server with the next-highest configured priority.

NAS is normally configured with:

- Max-retries: maximum number of times system will attempt to retry communications with a server before system fails over to a backup RADIUS server.
- RADIUS timeout: how long system will wait for a response from a RADIUS server before re-transmitting the request.
- Detect dead after 'x' consecutive failure: if AAA server is unreachable consecutive number of times then mark it as a dead server.

When RADIUS server does not respond to the request from NAS, NAS retransmits the same request to the same server until max-retries are exhausted. NAS marks that server as unreachable and tries to send the request to another server if configured. The new session request again goes to the same 1st server; retries until max-retries exhaust. In this process, if detect dead after 'x' consecutive failure exhaust, then NAS marks this server as dead.

Here, we assume RADIUS server as a remote server only and not as a Proxy server and also a system comprising of multiple RADIUS servers where selection of a server is based on round-robin algorithm.

Normally, in a dense area where there are more number of customers sending Create PDP Context request message to GGSN, multiple radius servers (more than 10) are used in round-robin fashion for load balancing.

II. PROPOSAL

With multiple RADIUS servers configured in the system, sending a keep-alive message is strongly discouraged, since it adds to load and harms scalability without providing any additional useful information [1]. When multiple servers are used in the network and if some of the servers go down (not responding), it is really time consuming to send request to each and every server if round-robin is used till it reaches the active server. With all the RADIUS requests going to multiple non-responsive servers, it also adds to load in the network. Figure 2 explains this.

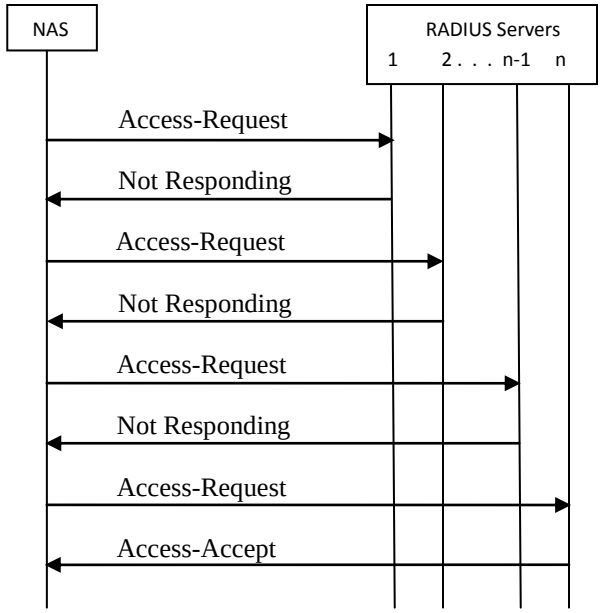


Figure 2. Non responding servers in Multi-Server System

Here, we propose to send an ACTIVE-Request from RADIUS server to NAS in case RADIUS server becomes active from dead state or becomes active for the first time. ACTIVE-Request will also add an attribute which will be its server ID. Also the new attribute values "Active Server ID" and "DEAD Server ID" for Type: 26 (Vendor-Specific) are introduced. This is based on [6].

12 Code = ACTIVE-Request
1 ID = 71
2 Length = 52
16 Request Authenticator

Attribute Type: 26 (Vendor-Specific)
Length: 12
Vendor Type: ACTIVE
Vendor Length: 6
Value: 00 00 01 Active Server ID

NAS will store the ID of the last active server and will update its priority table. So, when new request comes, NAS will look into its priority table and selects the active server in round-robin fashion. When some of the servers do not respond in a system of multiple RADIUS servers, with this new mechanism, NAS will select the appropriate active server first as it stores ID of the last active server and update the priority table. So, in meantime, if any of the non-responsive servers becomes active, it will send ACTIVE-Request to NAS. Then NAS comes to know about new active server, stores its ID and then sends the next new session request to this newly active server if its priority is less than the last active server else it will be selected in normal round-robin fashion. This way it will rather reduce

 ISBN: 978-1-61208-203-5

the load in the network by avoiding the retransmissions to the non-responsive server.

Algorithm:

If last Active Server received ID < Last Active Server
Last Active Server received ID = = Last Active server
Then follow the round-robin
Else place last Active Server received ID in round-robin queue matching priority and follow round-robin

The value of the attribute will be proprietary which will contain information about the last active server. So, NAS will select the active server for next session request to be sent based on this information and comparing the round-robin priorities.

After server sends ACTIVE-Request, NAS will send new request to this active server rather than trying to send requests to non-responsive servers. In Figure 3, server (n-1) sends ACTIVE-Request. So, NAS sends a new request to (n-1)th server directly; thus saving on requests and retransmissions of (n-2) servers considering all other servers before (n-1) are non-responsive.

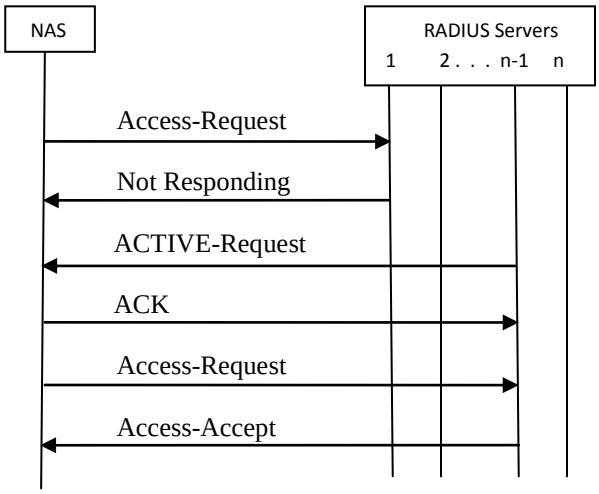


Figure 3. Server sends ACTIVE-Request

Let us assume following configuration at NAS:

Radius round-robin algorithm
Max-retries: 5
RADIUS timeout: 3 seconds
Detect dead after consecutive failure: 4
RADIUS Servers configured: 12
RADIUS dead time: 10 minutes
Keep alive not configured

In normal scenario, NAS will send request to 1st top priority server and that server will respond; 2nd request will go to 2nd server and so on.

Now, assume that all the servers are not responding except 12th server. So, as per above configuration, NAS will keep on sending request to 1st server until max-reties get exhausted. As radius timeout is 3 seconds, for each server radius request will take 18 seconds ((1+5) retries * 3 sec timeout).

For 1st request to reach 12th server, it will take (18 sec * 11 servers) = 198 seconds (3 minutes 18 seconds).
For 2nd request to reach 12th server, it will take (18 sec * 10 servers) = 180 seconds (3 minutes 0 seconds) and so on.

As detect dead after consecutive failure is 4, it will take 198 * 4 = 792 seconds = 13 min 12 seconds, before it marks 1st server as dead. After this, any request that comes from NAS will not be sent to 1st server but to 2nd server. As 2nd server also does not respond, this cycle repeats till it marks 2nd server as dead. This time it will take (198 -18) * 4 = 720 seconds = 12 minutes.

In this way, NAS will mark all servers dead except 12th one. As NAS is configured with no keep alive, NAS will never try to send any request to 1st server even if it is active or to any other server which is active until RADIUS dead time expires (which is 10 minutes in this case).

If, for example, 10th server becomes active, with new proposal, it will send ACTIVE-Request. NAS will then select this active server to send any new session request, thus saving on retransmissions. In meantime, if another server (e.g., 2nd server) becomes active, it will also send ACTIVE-Request. Then NAS will compare its priority table to select the appropriate active server for new request.

Selection of the active server can also be achieved by reducing the number of requests and retransmissions within a system of multiple RADIUS servers. But, this will not give fair amount of time to a particular server if that server becomes active before retransmissions are exhausted. Then, after retransmissions to non-responsive server are exhausted, there will be another round of retransmissions to the next selected server if that is also non-responsive. This will again add load in the system of multiple servers, even if number of requests and retransmissions are less.

We also propose to send DEAD-Request when the RADIUS server goes down. But, it has some limitation. In some scenarios like power outage or kernel panic, server does not get any chance to send any information.

13 Code = DEAD-Request
1 ID = 71
2 Length = 52
16 Request Authenticator

Attribute Type: 26 (Vendor-Specific)
Length: 12
Vendor Type: DEAD
Vendor Length: 6
Value: 00 00 01 DEAD Server ID

 ISBN: 978-1-61208-203-5

After receiving DEAD-Request, NAS will update its priority table and then selects the next available active RADIUS server for the new request. In this case, NAS will not retransmit the request to the dead server. If this dead server again becomes active, it again sends ACTIVE-Request and cycle repeats. Figure 4 shows this. In Figure 4, assume that server 2 is non-responsive and it sends DEAD-Request. Then NAS will send new request to next active server after 2nd, i.e., (n-1)th server, by looking into its priority table thereby saving on requests and retransmissions to 2nd server.

If 2nd server again becomes active, it will send ACTIVE-Request to NAS. As 2nd server's ID is less than the (n-1)th server, NAS will send new request to 2nd server instead of (n-1)th server.

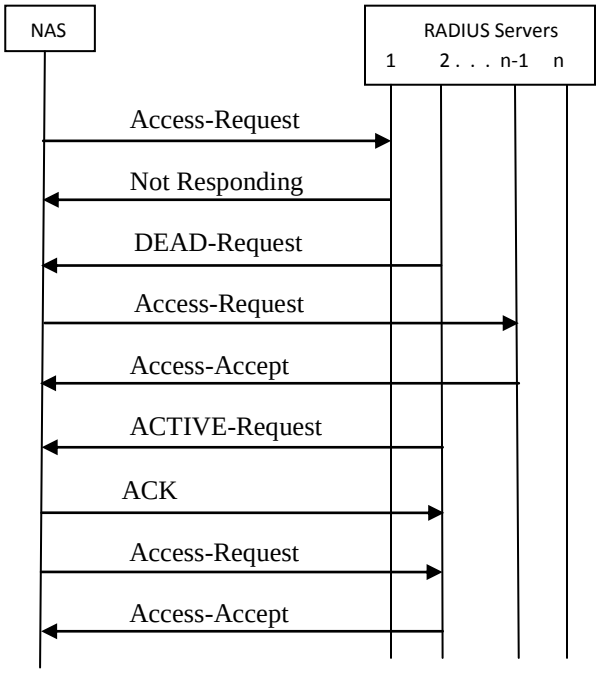


Figure 4. Server sends DEAD-Request

If we reduce the RADIUS dead time to minimum value (around 1 to 2 minutes), then NAS will send new requests to all dead servers in round-robin after this timer expiry and if it does not get any response then again marking of server dead cycle will start.

III. CONCLUSION AND FUTURE WORK

This proposal attempts to improve the communication efficiency between NAS and RADIUS server. It does so by allowing the RADIUS server to communicate its state (active/dead) to NAS. In some circumstances, the server does not get an opportunity to send anything when it goes offline like in case of power outage or kernel panic. But this proposal will help in effectively selecting RADIUS server.

Future work of this proposal includes simulation model of interaction between NAS and RADIUS server. It will also include the performance evaluation to support the concept in real system.

This proposal will help to select active server properly by avoiding retransmissions to the non-responsive servers thereby causing less CPU utilization in the network.

REFERENCES

[1] C. Rigney, S. Willens, A. Rubens, and W. Simspson, "Remote Authentication Dial In User Service (RADIUS)", IETF RFC 2865, June 2000.

[2] M. Chiba, G. Dommety, M. Eklund, D. Mitton, and B. Aboba, "Dynamic Authorization Extensions to Remote Authentication Dial In User Service (RADIUS)", IETF RFC 3576, July 2003.

[3] 3GPP TS 29.061 V9.2.0 (2010-03), page 56.

[4] A. DeKok, "Use of Status-Server Packets in the Remote Authentication Dial In User Service (RADIUS) Protocol", IETF RFC 5997, August 2010.

[5] 3GPP TS 32.295 V8.1.0 (2009-09), pp. 21-24.

[6] D. Mitton, "Network Access Servers Requirements: Extended RADIUS Practices", IETF RFC 2882, July 2000.

 ISBN: 978-1-61208-203-5

Attribute-based Group Key Management for Wireless Sensor Networks

A Cross-layer Design Approach for Group Key Management

Jun Noda
Cloud System Research Laboratories
NEC Corporation
1753 Shimonumabe, Nakahara-ku, Kawasaki 211-8666 JAPAN
Email: j-noda@cw.jp.nec.com

Yuichi Kaji
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0101 JAPAN
Email: kaji@is.naist.jp

***Abstract*—In this study, we investigated a group key management scheme that is especially suitable for large-scale wireless sensor networks (WSNs). Practical large-scale WSNs typically contain multiple groups of nodes, and the managing server needs to keep a number of group keys secure against possible attacks. We have developed a flexible and versatile formalization of an attribute-based group structure. The proposed formalization can model practical groups in real applications and enables secure and efficient management of multiple group keys. One of the key advantages of this approach is that a certain cross-layer design mechanism can be implemented in the group key revocation protocol. We show through computer simulation that adding some controlled redundancy in the upper-layer protocol helps reduce the lower-layer traffic in a realistic setting. The results demonstrate that using attribute-based groups is more suitable and practical for WSNs than the conventional group key management mechanisms.**

Keywords*-*wireless sensor network; security; group key; cross-layer design.

I. Introduction

Security is a crucial issue in many of the applications used in wireless sensor networks (WSNs). In this study, we focus on providing security through the management of cryptographic group keys owned by sensor nodes.

In a practical WSN with a large number of sensor nodes, nodes are typically sorted on the basis of their attributes, and the resulting *groups* of nodes are organized in the network. A group can function as a unit for the access control of critical information, so it is convenient if all nodes in a group are provided with an identical *group key* that is used to encrypt or authenticate critical data. The group key must be managed in such a way that it is known to group members only. When a node becomes compromised and is removed from the group, we need to replace the group key to prevent the removed node from accessing critical information. In this procedure, called *key revocation*, the key managing *server* selects a new group key and delivers it to all nodes that remain in the group.

Secure and efficient schemes for group key revocation have long been studied in the research arm of the information security field (see [2][6] for short surveys on the key findings of these studies). Unfortunately, most of these schemes are too complicated for WSNs, excepting a few simple schemes such as the LKH protocol [7]. Efforts have also been made to establish simple, lightweight mechanisms that explicitly focus on WSNs. For example, a simple unicast-based scheme can be constructed over SPINS [5]; LEAP [8] makes use of an internal timer to destroy critical information in each node, and [1] applies the idea of self-healing key distribution to WSNs.

Despite the insight gained from these previous works, we feel that some of the important aspects of group key management in WSNs have not been sufficiently considered thus far. For example, these other studies often assume that there is only one group in the network and therefore focus on the management of just one group key. In our view, this is fatally misleading approach. In a practical WSN, there are many groups in the network, and each group is typically managed by a single server. This means that the server can make use of some of the "safe" group keys in the revocation procedure of the "threatened" group keys. Such an organic use of multiple group keys helps make the key revocation much more efficient than managing multiple group keys separately and independently. Having multiple groups in the network is advantageous in terms of realizing secure and efficient key management, but investigation in this direction has not been considered so far. Another aspect we need to consider is the cross-layer design approach for the key management scheme. Previous studies assume that the communication in the lower layer is reliable, and the upper-layer protocols are designed so that there is no redundancy in the transmission of data packets. In the replacement of a group key, a node is not allowed to drop any of the data packets that are transmitted from the server to that node. This suggests that the retransmission of data packets will be requested everywhere across the network until all nodes receive all the required data packets through the unreliable wireless communication channel. If we add controlled redundancy to the upper-layer protocol, the issue of the retransmission of data can be completely mitigated.

The purpose of this paper is to refine and evaluate the

 ISBN: 978-1-61208-203-5

group key management scheme previously proposed by Noda et al. [4] with a focus on the above observation. The proposed scheme consists of two components: an *attribute-based group structure* and a *group key revocation protocol*. The attribute-based group structure is a mathematical formalization of the family of groups in WSNs. The formalization is so flexible that it can model the wide variety of groups that are typically present in a WSN. The key revocation protocol is what controls the replacement of a group key. In basic terms, the server broadcasts the encryption of information that is needed to update the group key. The key point here is that the encrypted messages are composed in such a way that every legitimate node has a chance to receive multiple messages from which it can learn the required information, and the key is replaced successfully even if some of these messages get lost during the communication. We have already outlined the overall concept of our scheme in a preliminary study [4], but there are still many points that must be refined and substantiated. In this study, we describe the proposed scheme in detail and evaluate its efficiency under realistic conditions. The protocol has a controllable parameter that changes the redundancy of the transmitted data, and computer simulation shows that having some redundancy in the upper-layer protocol helps reduce the total amount of communication traffic.

II. Related Work

There have been quite a few studies that focus on the management of group keys, but not all of them can be used in WSNs. For example, there have been studies that exploit the flexible properties of public-key cryptography, but it is still arguable if public-key cryptography is acceptable for sensor nodes with limited resources. The self-healing mechanism has been considered for WSNs in [1], but it remains unclear if the computation over a finite field with a large order is feasible for sensor nodes.

In the following, we restrict ourselves to those schemes that are based on lightweight symmetric-key cryptography. A conventional work that conforms to this condition is widely known as the LKH scheme [7]. In this scheme, we consider a tree-like structure in which nodes are attached with key-managing keys (KEK) and in which leaves correspond to group members. The KEK at the root node plays the role of group key. The revocation of the group key is performed by constructing encrypted messages based on the tree structure. If there are n members in the group, the server broadcasts $O(\log n)$ different messages. There are many variations and extensions of the LKH scheme, but perhaps [3] is the most significant. In [3], we consider a scenario in which there are multiple groups in the network, and a node (user) belongs to one or more groups simultaneously. The mechanism in [3] mitigates the overhead for managing KEK, but the functionality of key revocation is degraded and we occasionally need to perform off-line reconstruction of key trees.

In the early days of WSN research, investigations were made to construct rudimentary but lightweight mechanisms for group key management. For example, in ZigBee [9], a *global-key* (network key) is embedded in all nodes in the network. This global-key can be regarded as the group key of a group that consists of all nodes, but we cannot use it as the group key of an "internal" group that contains only some of the nodes in the network. We should also point out that there is no explicit mechanism that helps revoke the global-key. We can use the node keys (master keys) of ZigBee and SPINS [5] to allow the server to send a group key to legitimate nodes, but such a protocol is essentially unicast-based and not efficient for large groups with many members. LEAP [8] is a powerful scheme that allows sensor nodes to form arbitrary groups. Its primary drawback is that all nodes in the network must have a precise timer and an apoptosis mechanism that diminishes critical information in the node, which seems to be an unrealistic expectation.

III. Attribute-Based Group

Generally speaking, a group is a set of nodes that have certain characteristics in common. Multiple groups can be defined in the network based on different characteristics, and a single node may belong to multiple groups in general. To establish a versatile model of such groups, we define a group structure in terms of *attributes* and *attribute values*.

An attribute is a characteristic that is associated with nodes. For example, "deploy location", "manufacturer", "type of equipped sensor", and "the most significant byte of the MAC address" are examples of attributes. For each attribute, a sensor node has a unique attribute value, where we assume that a special "undefined" attribute value is allowed if the attribute is not affiliated with a particular group of nodes. A group can be regarded as a set of nodes that have the same attribute values for certain attributes.

We can provide a mathematical formalization of the above intuitive definition. Let N be the set of all nodes in WSN.

Definition 3.1: An *attribute* is a set partition $A = \{G_1, \ldots, G_m\}$ of N, that is, m is a positive integer, $G_j \subset N$ for $1 \le j \le m$, $G_{j_1} \cap G_{j_2} = \emptyset$ for $j_1 \ne j_2$, and $G_1 \cup \cdots \cup G_m = N$.

We intend G_j to be the set of nodes that have the j-th attribute value for the considered attribute. Assume that there are d different attributes $A_1, \ldots, A_d$ in the network. We write $A_i = \{G_{i,1}, \ldots, G_{i,m_i}\}$ for $1 \le i \le d$, where m_i is the number of attribute values of the i-th attribute. The set of nodes $G_{i,j}$ with $1 \le i \le d$ and $1 \le j \le m_i$ is called a *base set*.

Definition 3.2: A set G of nodes is called an *attribute-based group* with *rank* r if G is defined as

$$G = G_{i_1,j_1} \cap \cdots \cap G_{i_r,j_r} \tag{1}$$

with $1 \le i_1 < \cdots i_r \le d$ and $1 \le j_c \le m_{i_c}$ for $1 \le c \le r$.

Note that base sets are attribute-based groups with rank 1.

 ISBN: 978-1-61208-203-5

In a practical WSN, a group is defined as a semantically meaningful set of nodes, while the attribute-based groups are sets of nodes that are defined mechanically from given attributes. This means that there can be many attribute-based groups that have little significance from a practical viewpoint. However, we strongly expect that semantically meaningful groups are also attribute-based groups if the attributes are chosen appropriately. For example, "the group of nodes deployed on the second floor" can be obtained by using "deploy location" as one of attributes and having "second floor" be one of the attribute values. It is possible to define four attributes corresponding to four bytes of IP(v4) addresses (with attribute values $\{0, \ldots, 255\}$), and we can define "the group of nodes that belong to sub-net 192.1.2.*" as an attribute-based group with rank 3. We can use these attributes to consider an attribute-based group such as "the set of nodes that are deployed on the second floor and whose least significant byte of IP addresses is 123" while ignoring the meaningless attribute-based groups. In a sense, the attribute-based groups are a super-class of semantically meaningful groups. Therefore, we refer to attribute-based groups as simply *groups* in the following discussion.

For the practicality of discussion, we consider one additional condition for attributes, and assume that this condition is satisfied henceforth.

Definition 3.3: The set of attributes $A_1, \ldots, A_d$ is *complete* if no group with rank d contains two or more nodes.

This condition assumes that no two nodes have completely the same set of attribute values. This is quite a reasonable assumption because sensor nodes in the real world are all different in nature. Indeed, we can easily transform an incomplete set of attributes to a complete one by introducing an additional attribute that is based on device-unique identities such as a MAC address or that plays the role of a sequence number in a group. We should also point out that completeness implies that, for each node $n \in N$, there exists an attribute-based group G, such that $G = \{n\}$. The group key of G, which we define in the next section, can be used like the node key of ZigBee [9] owned by n and the server.

Example 3.1: Figure 1 shows an example of a complete set of attributes, where $N = \{n_1, \ldots, n_7\}$, and

$$G_{1,1} = \{n_1, n_2, n_3\}, G_{1,2} = \{n_4, n_5\}, G_{1,3} = \{n_6, n_7\},$$
$$G_{2,1} = \{n_1, n_4\}, G_{2,2} = \{n_2, n_5, n_6\}, G_{2,3} = \{n_3, n_7\},$$
$$G_{3,1} = \{n_1\}, G_{3,2} = \{n_2, n_4\}, G_{3,3} = \{n_3, n_5, n_6, n_7\}.$$

In the key revocation procedure, which is discussed in later sections, we need to consider a *set cover* of nodes. The set cover problem is a classical subject in computer science, but the set cover we need here is slightly different from the conventional one.

Definition 3.4: Let m be a positive integer and $n \in N$. A collection of base sets $\mathcal{G} = \{G_1, \ldots, G_l\}$ is an (m, n)-*cover* if the following conditions hold.

1) For each $n' \in N \setminus \{n\}$, $\mathcal{G}$ contains m or more base sets to which n' belongs.
2) For any i with $1 \le i \le l$, $n \notin G_i$.

Note that every node except n is covered at least m times by base sets in $\mathcal{G}$. The parameter m is called the *multiplicity* of the cover $\mathcal{G}$. Consider the attributes and groups that are given in Example 3.1. The collection of base sets $\mathcal{G} = \{G_{1,2}, G_{2,2}, G_{2,3}, G_{3,2}, G_{3,3}\}$ is a $(2, n_1)$-cover. For example, the node n_3 has two base sets $G_{2,3}$ and $G_{3,3}$ in $\mathcal{G}$ to which n_3 belongs. In the collection $\mathcal{G}$, there is no base set that contains n_1.

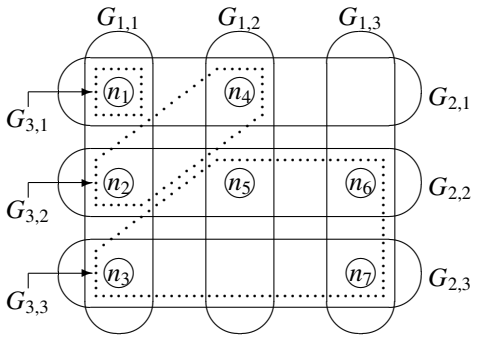


Figure 1. An example of a complete set of attributes.

IV. Group Keys and Revocation Procedure

We have investigated secure and efficient group key management for attribute-based groups. In our framework, elementary information is distributed to sensor nodes, and group keys are then computed from this information.

A. Assumptions

First, we clarify the security assumption that is needed in the investigated framework.

The first assumption we need is that the compromise of a node is promptly detected by the server, which immediately reacts to revoke the compromised node. This condition is quite feasible in many WSN systems in which sensor nodes are watched by somebody, or equipped with a small physical security mechanism, and hence the condition does not interfere with the practicality of the procedure. We remark that this condition implicitly assumes that a malicious attacker does not compromise multiple nodes at one time, which is essential in the following protocol. We also assume that all nodes agree with a cryptographic hash function h and a symmetric-key cryptography. We write $E(k, x)$ to denote the result of the encryption of x using k as a key.

Each node has two tables of keys, called an *active-key table* and a *next-key table*. The former is used to store cryptographic keys that are currently active and used in the network, and the latter is used to store keys that will be used when the currently used keys expire. The active-key table is associated with the *version number* and the *expiration time*,

 ISBN: 978-1-61208-203-5

and the next-key table is associated with the *version number*, the *activation time*, and the *expiration time*. The usage of these information will be explained later.

B. Group Keys

Assume that each base set is associated with secret information, which is called *base key*. The base key of $G_{i,j} \in A_i$ with $1 \le i \le d$ and $1 \le j \le m_i$ is denoted by $k_{i,j}$. The base keys are managed in such a way that a node knows $k_{i,j}$ if and only if that node belongs to the base set $G_{i,j}$. Group keys are defined by using base keys as follows.

Definition 4.1: Let G be an attribute-based group defined as (1). The *group key* of the group G is $k(G) = h(k_{i_1,j_1} \| \cdots \| k_{i_r,j_r})$, where "$\|$" is the concatenation of keys.

It is easily understood that a node is able to compute $k(G)$ if and only if that node belongs to the group G.

At the system initialization, the server determines a base key $k_{i,j}$ for each $1 \le i \le d$ and $1 \le j \le m_i$. The server also determines the *version number* and the *expiration time* of this set of base keys. The server delivers $k_{i,j}$ to all nodes in $G_{i,j}$ through a secure communication channel, together with the version number and the expiration time. The delivery of the base key and related information is performed in a safe place, possibly before nodes are deployed.

A node receives base keys and related information from the server and records them to the active-key table. At this time, the next-key table is empty. The key tables are managed by each node according to two principles:

- The content of the next-key table overwrites that of the active-key table when it gets to the activation time of the next-key table.
- If the active-key table is going to expire but the content of the next-key table has not been received yet, the node sends a NACK message to the server to request (re)transmission of the key information.

The server activates and expires base keys so that the keys are synchronized between the server and nodes. The server also receives NACK messages from sensor nodes and responds to nodes by sending requested information in a unicast manner, where the information is encrypted by using a group key that is available to the destination node.

C. Key Revocation

Group keys are replaced for two primary reasons. The first is that the keys tend to be used for too long a time. In general, it is not recommended to use a single key for very long because this gives attackers an opportunity to make a cryptanalysis and increases the risk of possible key leakage. From the security viewpoint, the periodical replacement of group keys is always recommended, even if there is no apparent security issue. The other reason for replacing a group key is the revocation of nodes. If a node in the network is compromised by somebody, we must consider a worst-case scenario, i.e., that an attacker has accessed all the information stored in the node. Cryptographic keys stored in the node are no longer secure, and we must replace them immediately. In such a case we need to deliver updated group keys to all nodes in the group except the compromised node. This procedure is called group key revocation. Generally speaking, the periodical replacement of group keys can be regarded as a special case of a group key revocation with no node revoked. In the rest of this section, we describe the group key revocation in detail.

Consider a scenario in which the server detects that a node $n \in N$ has been compromised by a malicious attacker. Without loss of generality, we assume that the node n belongs to d base sets $G_{1,1}, \ldots, G_{d,1}$ and has d base keys $k_{1,1}, \ldots, k_{d,1}$. We need to replace these d base keys because the attacker may discover them by disassembling the node n. The straightforward approach to solving this issue is to deliver a new base key (a replacement for $k_{i,1}$) to legitimate nodes in $G_{i,1} \setminus \{n\}$ for each of $1 \le i \le d$. However, to simplify the communication and to devise the cross-layer mechanism of the protocol, we consider a protocol in which the server sends a single "modifier" to all nodes other than n and enables nodes to replace their own base keys by using the modifier information. The modifier must be protected by encryption in such a way that it is accessible from all nodes other than n and that node n cannot learn what the modifier is. To achieve this requirement, the server performs the following procedure.

1) Determine the multiplicity parameter m and modifier string s. Also determine the version number v', the activation time a', and the expiration time e' of the updated set of base keys. The version number v' must be bigger than the version number of the currently used base keys.
2) Compute an (m, n)-cover $\mathcal{G} = \{G_{i_1,j_1}, \ldots, G_{i_l,j_l}\}$, where $1 \le i_c \le d$ and $1 \le j_c \le m_{i_c}$ for $1 \le c \le l$.
3) Broadcast l messages

$$M_c = (i_c, j_c, E(k_{i_c,j_c}, s \| v' \| a' \| e')) \qquad (1 \le c \le l). \quad (2)$$

Upon receiving the message M_c, a node n' performs the following procedure.

1) Discard the message if $n' \notin G_{i_c,j_c}$. Otherwise (i.e., $n' \in G_{i_c,j_c}$), proceed to the next step.
2) Decrypt the third component of M_c and retrieve s, v', a', and e'.
3) Do either one of the following operations.
   - If the next-key table is not defined, then record $h(k_{i,j} \oplus s)$ as the next base key of the base set $G_{i,j}$ with $n' \in G_{i,j}$ (and hence n' knows $k_{i,j}$),
   - If the next-key table is defined but its version number is smaller than v', discard the content of the table and perform the above operation.
   - In other cases, the next-key table is not modified.

 ISBN: 978-1-61208-203-5

In updating key information in the step 3, the hash function h is used to mitigate the risk of possible leakage of old keys. Without this hash function, an adversary who happens to know old keys may find the latest keys by investigating the XOR relation between new and old keys. Remark also that the update of the next-key table is controlled by the version number. This is to avoid possible confusion caused by the delay of message delivery, duplicated delivery of the same message, replay attacks by adversaries, and so on. As we explained previously, the next-key table replaces the active-key table when the appropriate time comes.

The focal point of the above protocol is that the messages M_c in (2) are determined from an (m, n)-cover. The server prepares messages in such a way that every legitimate node is given m or more messages that make sense to that particular node. In wireless communication, we cannot avoid the fact that some of these messages will be lost during the communication; however, we do not have to be too nervous about this because if just one of the m messages is delivered to a node, that node can successfully update its base keys. The multiplicity m controls the redundancy of transmitted messages, and the redundancy mitigates the effect of communication failure in the lower layer of communication.

V. Experiment

We performed preliminary experiments to determine how the multiplicity of a cover affects the total amount of traffic in a realistic setting. These experiments were preliminary in two senses. First, we consider the periodical replacement of group keys and ignore the node revocation scenario. This is because we need to introduce many additional parameters and assumptions if we would like to consider compromised nodes. The second reason we consider these experiments preliminary is that we do not discuss the network-wide burden of the protocol. We will evaluate two significant quantities related to the protocol, but will not consider other aspects of the protocol or the network. In this sense the experiment is limited, but this simple setting is effective in terms of concentrating on the cross-layer effect of multiplicity on the actual traffic.

We consider a WSN that contains 1,024 nodes with one of the nodes playing the role of the server. The nodes are deployed so that they form a square grid matrix of 32 $\times$ 32 nodes. The dimension of one unit grid is 8 m per side, meaning that the 1,024 nodes are deployed in a 248 $\times$ 248 m field. We assume that the node at the (x, y) position $(0 \le x, y \le 31)$ is given a *sequence number* $x + 32y$. The node with the sequence number 512 is deployed at almost the exact center of the field, and it plays the role of the designated server. We define ten attributes $A_1, \ldots, A_{10}$ with $A_i = \{G_{i,0}, G_{i,1}\}$ for $1 \le i \le 10$ in such a way that a node with a sequence number n belongs to $G_{i,j}$ if and only if the i-th most significant bit of the binary representation of the number n equals j. For example, the node with the number $858 = (1101011010)_2$ belongs to

$$G_{1,1}, G_{2,1}, G_{3,0}, G_{4,1}, G_{5,0}, G_{6,1}, G_{7,1}, G_{8,0}, G_{9,1}, G_{10,0}.$$

The nodes are placed in a grid matrix manner, so it is natural to consider that nodes in a row, or in a column, consist of one group, and that the network contains 64 groups in total. Such a group can be represented as an attribute-based group. For example, the group of nodes in the first row (i.e., $y = 0$) is defined as an attribute-based group of rank five $G_{1,0} \cap \cdots \cap G_{5,0}$, which contains nodes with numbers from $0 = (0000000000)_2$ to $31 = (0000011111)_2$.

In the preliminary experiment, we investigated the traffic when replacing the group keys for these 64 groups. We used QualNet for the computer simulation. The assumed protocols were IPv4 in the network layer and IEEE 802.15.4 in the MAC and PHY layer. The physical layer payload was 127 bytes, which is sufficient to contain the messages M_c in (2) in one packet. The wireless communication used a 2.4-GHz band with O-QPSK modulation. The communication speed was 250 kbps. The transmission power, antenna gain, and related parameters were adjusted so that the wireless range was almost equal to 14 m. Specifically, the transmission power was -17 dbm and the antenna gain was set to -3.0 dB. We employed the free space propagation model. The 14-m range allowed a node to communicate with eight neighbor nodes. A routing tree was manually provided, and the server and each node communicated with each other in a multi-hop manner. The server took 200 ms for the transmission time interval of packets, and nodes tried to avoid possible collision by using random jitter, where the jitter time was upper-bounded by various constants. For the control of broadcast messages in the network layer, nodes inspected the communication of their children with the passive ACK principle in which the waiting time was 150 ms. Retransmission of packets in the network layer was permitted up to three times.

At the beginning of the protocol, we set the key tables of nodes so that the active base keys expire in 60 seconds while the next-key table is empty. The server transmits messages for the periodical replacement of base keys and the nodes process the received information. After 60 seconds, nodes that could not obtain new base keys start sending NACK messages. A node sends one NACK message every second until it succeeds in updating its node keys. Upon receiving NACK messages, the server retransmits the modifier information to individual nodes that have sent NACK messages. The duration of the simulation was 120 seconds, which is sufficient for all nodes to finish replacing their base keys.

Figure 2 shows the number of packets that are transmitted by the server for the sake of key replacement, including packets transmitted as responses to NACK messages. The x-axis of the graph is the multiplicity m of the cover, and the y-axis is the number of packets. In the periodical replacement of group keys, the server initially transmits $2m$ packets,

 ISBN: 978-1-61208-203-5

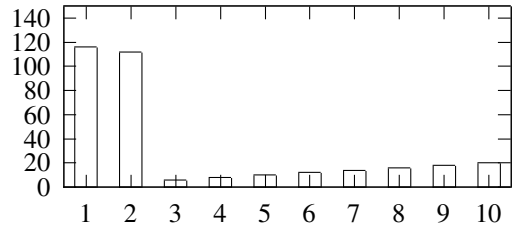


Figure 2. The number of packets transmitted by the server.

Table I
THE NUMBER OF NODES THAT ISSUE NACK MESSAGES.

Multiplicity m	NACK nodes	NACK messages
1	13	396
2	6	289
3–10	0	0

one for each $G_{i,j}$ with $1 \leq i \leq m$ and $j = 0, 1$. If every node receives one or more packets from which it can obtain modifier information, the retransmission of packets is not necessary. Such a favorable scenario is more likely if the multiplicity m is large, because large multiplicity means that nodes are given a greater chance of finding the modifier information. Indeed, the graph shows that the retransmission of packets is not needed for $m \geq 3$. On the other hand, if m is a small value, then some nodes fail to receive the required information. Such nodes send NACK messages to the server, and the server needs to send additional packets as responses to NACK messages. The graph clearly shows that having redundancy with controlled multiplicity helps decrease the total number of packets transmitted by the server.

Table I shows the other aspect of the protocol by showing the number of nodes that needed to issue NACK messages and the total number of NACK messages issued by nodes. If m was small, some nodes failed to receive the initial packets and had to send NACK messages to the server. We can see that one node issued several NACK messages, which implies that the NACK processing takes a rather long time. Detailed observation suggests that if a packet is lost near the server, many subsidiary nodes send NACK messages, which results in an unfavorable increase of communication traffic.

As above, adding some redundancy is good for reducing the total communication traffic in a realistic setting. We remark however that adding redundancy is not always as easy and efficient as in the attribute-based key management case. For comparison, consider that the LKH scheme is used to manage 64 group keys in the above experiment. In an idealized environment, the server sends 64 packets for periodical key replacement because only one packet is needed for each key tree. The observation in the above experiment suggests, however, that there will be 13 events in which a node cannot receive a packet. With 64 transmitted packets, the total number of such events is estimated to 832, and a large number of retransmission will be requested. The number of such events may be reduced by sending one packet m times, though, it increases the number of initially transmitted packets to $64m$. Remind that the number of initially transmitted packets is $2m$ in the attribute-based key management approach, and therefore, the overhead for increasing the multiplicity is more in LKH than the attribute-based key management.

VI. CONCLUSION

The cross-layer design effect in a group key management is discussed. It is shown that having controlled redundancy contributes to reduce the total amount of communication traffic in a realistic setting. This effect may exist in any key management protocols, but we saw that the overhead for having redundancy in the attribute-based key management scheme is smaller than that for widely known LKH scheme. From these results and observations, the attribute-based key management is concluded to be suitable for managing multiple group keys in large and practical WSNs.

ACKNOWLEDGMENTS

This work was partly supported by the Ministry of Internal Affairs and Communications (MIC).

REFERENCES

[1] R. Dutta, E. Chang, and S. Mukhopadhyay, Efficient Self-Healing Key Distribution with Revocation for Wireless Sensor Networks Using One Way Key Chains, 2007 Applied Cryptography and Network Security, pp. 385–400, 2007.

[2] B. Jiang and X. Hu, A Survey of Group Key Management, 2008 Intl. Conf. on Computer Science and Software Eng., Wuhan, China, pp. 994–1002, 2008.

[3] E. Jung, A. Liu, and M. Gouda, Key Bundle and Parcels: Secure Communication in Many Groups, Computer Networks, 50, pp. 1782–1798, 2006.

[4] J. Noda, Y. Kaji, and T. Nakao, A Group Key Management Scheme for Sensor Nodes Belonging to Multiple Large-Scale Groups, Journal of Information Processing, *52*, 3, pp. 1160-1172, 2011 (in Japanese).

[5] A. Perrig, R. Szewczyk, J.D. Tygar, V. Wen, and D.E. Culler, SPINS: Security Protocols for Sensor Networks, Wireless Networks, *8*, 5, pp. 521–534, 2002.

[6] S. Rafaeli and D. Hutchison, A Survey of Key Management for Secure Group Communication, ACM Computing Surveys, *35*, 3, pp. 309–329, 2003.

[7] C.K. Wong, M. Gouda, and S.S. Lam, Secure Group Communications Using Key Graphs, 1998 ACM SIGCOMM Conf. on Applications, Technologies, Architectures, and Protocols for Comput. Comm., pp. 68–79, 1998.

[8] S. Zhu, S. Setia, and S. Jajodia, LEAP: Efficient Security Mechanism for Large-Scale Distributed Sensor Networks, 10th ACM Conf. on Comput. and Comm. Security, pp. 62–72, 2003.

[9] IEEE 802.15 WPAN Task Group 4 (TG4), http://www.ieee802.org/15/pub/TG4.html ⟨10.04.2012⟩

 ISBN: 978-1-61208-203-5

Location-Based Utilization for Unidirectional Links in MANETs

Huda AlAamri and Farzad Safaei
ICTR, University of Wollongong, NSW 2522, Australia
E-mails: hmaaa634, farzad @uowmail.edu.au

Mehran Abolhasan and Daniel Franklin
University of Technology Sydney, NSW 2007, Australia
E-mail: Mehran.Abolhasan, Daniel.Franklin @uts.edu.au

***Abstract*—Heterogeneous Mobile Ad hoc Network(HMANET) comprises different nodes with different capabilities. Hence, transmission and receiving capabilities are different. This causes unidirectionality problem. Avoidances is the most used strategy in researches to route data, e.g., Blacklist. In this paper, we proposed a strategy for on-demand routing protocols to detect unidirectional link and resolve it in timely fashion. This strategy is based on utilizing locations of nodes to filter and cache incoming RREQ packets to find reliable path to destination in the existence of unidirectional links. Simulation results show that our strategy outperforms Blacklist strategy in homogeneous and heterogeneous MANET.**

***Keywords*-MANET; routing protocol; unidirectional link; AODV;**

I. INTRODUCTION

Mobile Ad hoc NETworks (MANETs) are networks of wireless mobile nodes that have no fixed structure. Each node may act either as a router or an end-user node. In MANETs, node heterogeneity is one of the main network conditions that significantly affects the performance of the routing protocols [1]. Although most current MANET routing protocols assume homogeneous networking conditions where all nodes have the same capabilities and resources, in real life MANET may consist of heterogeneous nodes that have different capabilities and resources like military (battlefield) networks and rescue operations systems. Hence, the transmission reachability and quality of data reception among nodes are different. This can create a problem of unidirectional link between any two nodes. Unidirectional link problem is defined, where node B has a higher transmission range than node A (see Figure 1). Therefore B includes A in its transmission range while A does not include B. Consequently, the link between B and A is unidirectional from B to A only. However, most reactive routing protocols in MANET, assumes all links between two nodes are bidirectional, which gives incorrect routing information. Therefore, this incorrect information creates large delay and packet loss in heterogeneous networking [1]. In [2], suggested that unidirectional link can be utilised to increase packet delivery and hence increase reliability. In this paper, we investigate this issue by proposing a strategy that is Location-Based Utilization (LBU) to detect and utilise unidirectional links in route discovery process of on-demand routing protocols. This strategy utilizes locations of forwarding nodes of RREQ packets to resolve unidirectionality problem. All received RREQ packets are cacheed and filtered before they are processed or dropped.

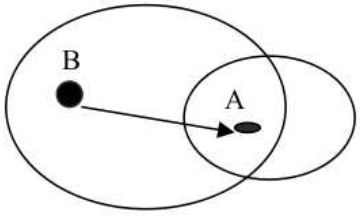


Figure 1: Unidirectional Link

The rest of this paper is organised as follows. Section II presents related works. In section III, the proposed strategy is described. The simulation parameters and scenarios that are used to investigate the performance of the proposed strategy are given in section IV. Then the results of the simulation study are summarised in section V. Section VI concludes the paper.

II. RELATED WORK

The common approaches to detect unidirectional link in MANETs are via MAC layer or network layer or both. Two way handshake Request-To-Send (RTS) and Clear-To-Send (CTS) is the common approach in MAC to avoid unidirectional links [3]. Network layer approaches use feedback mechanism either to detect and avoid unidirectional link or utilise it to improve routing processes. In heterogeneous MANET (HMANET), the issue of unidirectional links has been investigated. Different strategies have been developed to enhance the performance of routing protocols in presence of unidirectional links [2][3][4][5][6][7]. In AODV-Blacklist [8], when destination node sends RREP (or any node relays RREP) to next hop in the reserve path, it waits for ACK of receiving RREP. If it fails to receive ACK because of unidirectional link , then next hop is cached in blacklist. This means that when the node receives RREQ for second time from the node in blacklist, the packet will be dropped. AODV-BlackList avoids unidirectional link but with cost of high load of control overheads. Also, delay is increasing because source node may consume all RREQ_RETRIES to find path to destination.

 ISBN: 978-1-61208-203-5

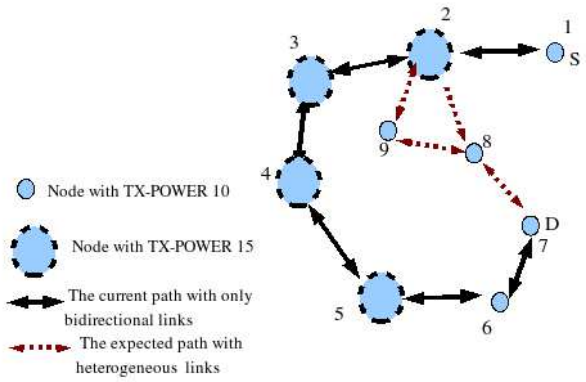


Figure 2: Routing through bidirectional links.

In [9], Early Unidirectionality Detection and Avoidance (EUDA) mechanism is proposed to detect and avoid unidirectional link in ad hoc network. This mechanism appends the forwarding node location only in RREQ packet to detect the unidirectionality. When node receives RREQ packet for the first time, it compares the transmission range to the distance to forwarding node using location information. If there is unidirectional link then the packet is dropped without any processing. This mechanism is used only to detect and avoid unidirectionality of links without utilizing it. In the worst case where there is no bidirectional route, all RREQ_RETRIES are consumed. Consequently, the control overheads increases and the packet delivery ratio decreases as the path to destination is not establishes. In [2], a powerful and simple strategy has been suggested to resolve unidirectional links in AODV-Blacklist. This strategy is developed to resolve the problem of unidirectional links by rebroadcasting RREP to first hop nodes as unidirectional link is detected and no nodes are blocked. To avoid insufficient exchanging ACKs during RREP rebroadcasting, TTL is set to 1. Also source node id and destination id are cached to avoid duplications of the same RREP packets. The simulation result shows improvement of AODV performance in term of packet delivery ratio and control overhead.

III. DESCRIPTION

In LBU, we use the concept of detecting unidirectional link using location information as in [9]. However, LBU differs from EUDA in [9] by utilizing the unidirectionality to improve routing process in on-demend routing protocols using 2 hops nodes locations . In Figure 2, there are different nodes with different transmission powers. Source node 1 initiates route discovery to find path to node 7. Node 2 will rebroadcasts the RREQ packet. Node 8 will receive the packet and has path to destination 7. However, it fails unicast its RREP to node 2 because of unidirectional link. As receiving a duplicated RREQ packet is ignored, then

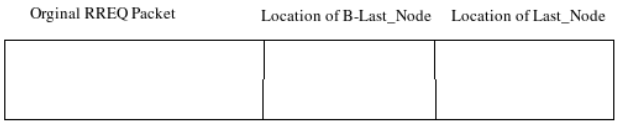


Figure 3: RREQ packet formate

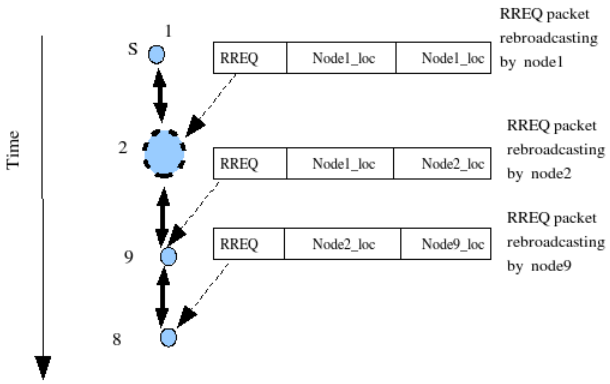


Figure 4: RREQ packet traversing

received RREQ from node 9 is ignored in node 8. In AODV-BlackList, rebroadcasting will continue until the destination is found or RREQ retries limit is reached. In Figure 2, node 8 will consider node 2 is unreachable and then inserts node 2 in its blacklist. Therefore, when node 8 receives any packet from node 2, it will be ignored. Source node 1 will have long path $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 7$ to reach destination 7, which 3-hop far. This long path can degrade the reliability of network and delay data comparing to expected path $1 \rightarrow 2 \rightarrow 9 \rightarrow 8 \rightarrow 7$. One of the strategy to resolve this problem, when node 8 detects unidirectional link to previous forwarding node of RREQ, it rebroadcasts its RREP to its first hop neighbours. As node 9 hears rebroadcasting of RREQ and RREP packets, it will unicast RREP packet to node 2. This idea is similar to [2] (see more details in related work section). However, this may create large number of paths and increases control overheads, which may degrade the network performance. Instead of rebroadcasting RREP as in [2], each node (e.g., node 8) starts caching all RREQ packets of the same source and flood id to resolve any unidirectional links. The description of how to detect and utilize unidirectional links are described in below subsections.

A. How is the unidirectionality detected?

Each RREQ packet will have two more fields, see Figure 3. These two fields carry locations of last two hops nodes, see Figure 4. When node receives RREQ packet,

1) The node calculates the distance to the forwarding

source id	flood id	forwarding node address	B_last_loc	last_loc	isUnidirLink

Figure 5: Seen_Data_Table formate

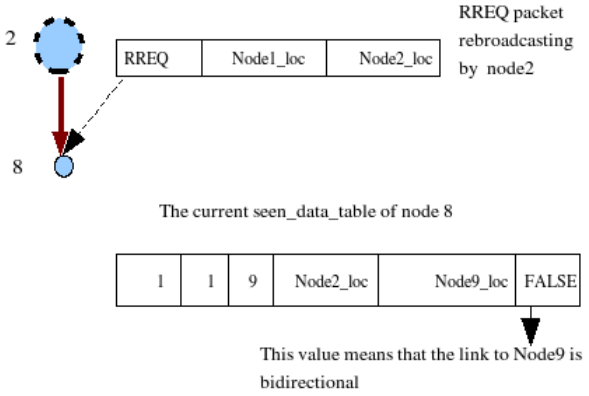


Figure 6: Incoming RREQ packet through unidirectional link

node by using locations information in RREQ packet.
$curr_{loc}$=(x_{curr}, y_{curr})
$lastloc$=$getLastLoc(Recived\_RREQ\_Packet)$
$Distance$=$\sqrt{(x_{curr} - x_{lastloc})^2 + (y_{curr} - y_{lastloc})^2}$

2) The node calculates its transmission range. In QualNet, the function PHY_PropagationRange calculates an estimated radio range for a given interface.

Transmission_Range_CurrNode=
PHY_PropagationRange(node, 0, FALSE)

3) The node compares its transmission range to the distance in step1:
if (Distance > Transmission_Range_CurrNode)
link is unidirectional
else
link is bidirectional

B. How is the unidirectional link fixed?

Unidirectional link here means that the current node can not reach the forwarding node while the forwarding node can reach it. Instead of dropping all RREQ packets of the same flood id and source id, the node caches the information in RREQ packet to detects and resolve the unidirectionality during the flooding of the same RREQ packet. These information is stored in a table called ”seen_data_table”,

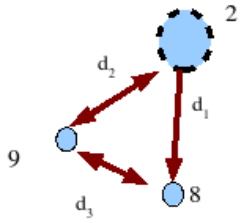


Figure 7: The triangle inequality in distance between nodes

which is similar to seenTable in AODV. seenTable is used to avoid duplication of the same RREQ packet where it keeps the id of source and flood number of the first incoming RREQ packet. seen_data_table is used to resolve unidirectionality. The formate of this table is shown in Figure 5. Each node receives RREQ packet, detects the type of the link. Each type of link has different process as following:

Unidirectional Link:

If the link is unidirectional (see Figure 6) then node searches seen_data_table for a record, which can resolve the problem where:

1) source id and flooding number are the same as the current received RREQ packet. This guarantee the freshens of nodes locations information and updates unidirectionality situation in timely fashion within neighbourhoods nodes.
2) The value of isUnidiLink is false, which means the forwarded node has bidirectional link to the current node.
3) Location value of B_last_loc field of the record is same as Location of Last_Node in received RREQ packet. In Figure 6, node 8 looks in its seen_data_table for a node that can reach the forwarding node of the current RREQ packet.
4) To avoid long path and replace unidirectional link with only 2-hop link, node is selected based on its location to form triangle inequality with current and forwarding node . In other words, we prefer the situation where the length of unidirectional link is less than the sum of lengths of other 2 links as shown in Figure 7 where $d1 < d2 + d3$ and di is the distance between node pair of nodes.

If a record is found that satisfies above conditions then the ”forwarding node address” in seen_data_table is used as next hop to the current forwarding node of the current received RREQ packet. Otherwise, information about the RREQ packet and unidirectionality are inserted in seen_data_table. Also if the received RREQ packet has not been process yet, then the packet will be processed after unidirectional link is fixed where node 9 will be the source of the packet. To utilise

 ISBN: 978-1-61208-203-5

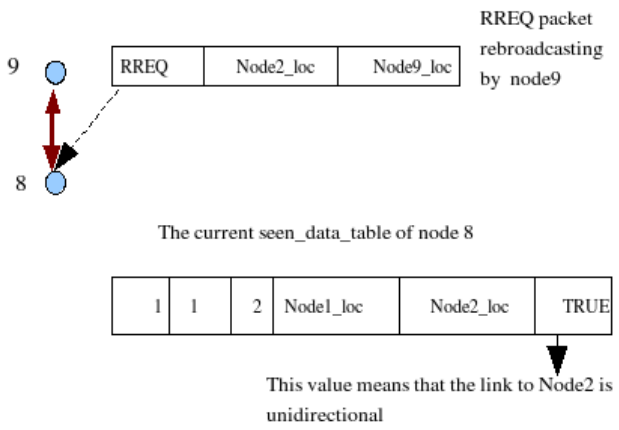


Figure 8: Incoming RREQ packet through bidirectional link

node memory, each record that have been used to solve unidirectionality in seen_data_table is deleted. Therefore, the record about node 9 in seen_data_table is deleted because it has already been used to solve the unidirectional link between node 8 and node 2. Consequently, as this problem has been solved, it is inefficient to insert information about node 2 in seen_data_table.

Bidirectional Link:

If the link between forwarding node of the current received RREQ packet is bidirectional (see Figure 8) then information of this packet is used to solve any unidirectional link in seen_data_table if:

1) Conditions 1 and 4 are satisfied as above described.
2) The value of isUnidiLink is true, which means the forwarded node has unidirectional link to the current node.
3) Location value of last_loc of the record is same as location of B_Last_Node in received RREQ packet. In Figure 8, node 8 looks for a node where the forwarding node of the current RREQ packet can reach it while node 8 can't.

If a record is found that satisfies above conditions then the address of current forwarding node is used as next hop to the forwarded node of the recorded packet. In Figure 8, node 9 will be the next hop to node 2. Otherwise, information about the RREQ packet and bidirectionality are inserted in seen_data_table to be used to solve any incoming unidirectional link. To utilise node memory, each record of the unidirectionality that has been solved seen_data_table is deleted. Therefore, the record about node 2 in seen_data_table is deleted. Each node receives second flood of the same RREQ packet will delete all records about the first flood in the seen_data_table.

IV. Simulation Models

The performance of LBU for unidirectional link is compared to BlackList and RTS/CTS strategy. AODV[8] and OTRP[10] are used as routing protocols. OTRP combines the idea of hop-by-hop routing such as AODV with an efficient route discovery algorithm called Tree-based Optimized Flooding (TOF) to improve scalability of Ad hoc networks when there is no previous knowledge about the destination. To achieve this in OTRP, route discovery overheads are minimized by selectively flooding the network through a limited set of nodes, referred to as branching-nodes. Those protocols have been simulated using the QualNet4.5 package. The simulations ran for 200s with 100 different values of seeds. Nodes density of 100 were randomly distributed on 1500 x 1500 grids. Random way point was used as mobility model with five different values of pause times that were 0s, 50s, 100s, 150s, and 200s. Speeds of the nodes were varied from 0 to 20 m/s. The simulated protocols have been evaluated with 10 data traffic flows. Constant Bite Rate (CBR) was used to generate data traffic at 4 packets per second. Each packet was 512 bytes. IEEE 802.11b was used as MAC protocol with constant transmission bandwidth of 2Mbps. The strategy has been evaluated in homogeneous and heterogeneous MANET where transmission power of all node was 15dbm in homogeneous MANET. In heterogeneous MANET, there are two different types of nodes where 50% of nodes have transmission powers as 15dbm and other 50% has transmission powers as 10dbm. Packet Delivery Ratio (PDR), End-to-End Delay, and Normalized Control Overhead (NCO) were used as performance metrics of each protocol. In addition, we introduce new metric called Retried Ratio (Ret_Ratio), which is ratio of the number of RREQ packets retried to the number of RREQ packets initiated. This ratio calculates the number of RREQ retries that has been consumed to find routes. As this ratio is high indicating that more RREQ retries have been consumed to find path to destination. Confidence interval of 95% is used to scale the data.

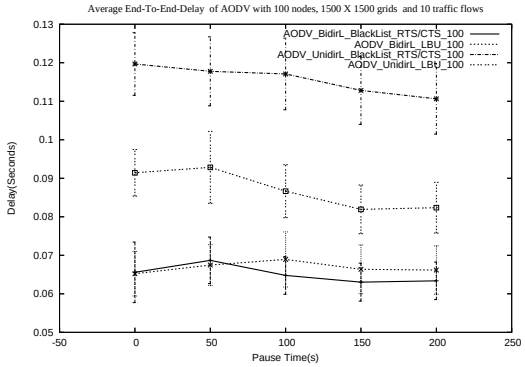
V. Results

LBU is applied on top of AODV and OTRP where Blacklist and RTS/CTS are disabled, see (Figure 9- Figure 10).

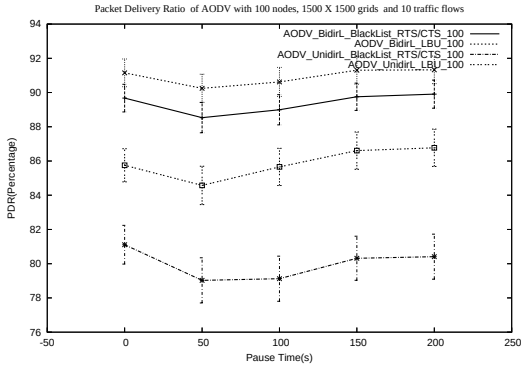
The problem of the unidirectionality affects routing process of on demand routing protocols, where the forwarding node of the RREQ may have unidirectional links to its neighbours nodes. In other words, rebroadcasting nodes stores incorrect information about the first hop, which is unreachable because of unidirectionality. Consequently, source node does not received RREP packet and then the route may not found. This will increase the number of route discovery occurrences and consequently increases Ret_Ratio.

Blacklist_RTS/CTS strategy with AODV and OTRP detect unidirectional links after it occurs then avoids unidirectional links without solving. This strategy may

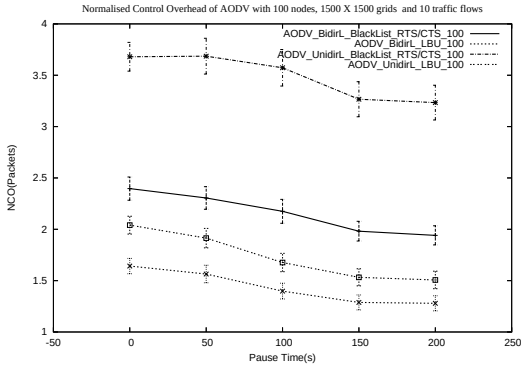
 ISBN: 978-1-61208-203-5



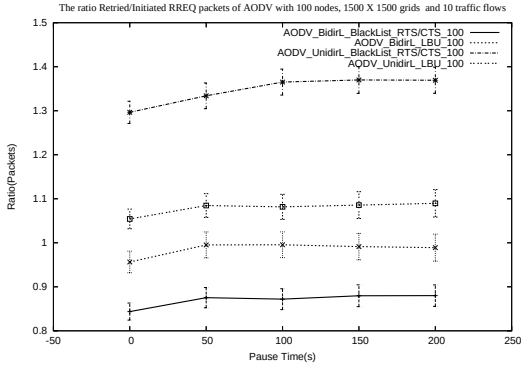
(a) Delay



(b) PDR

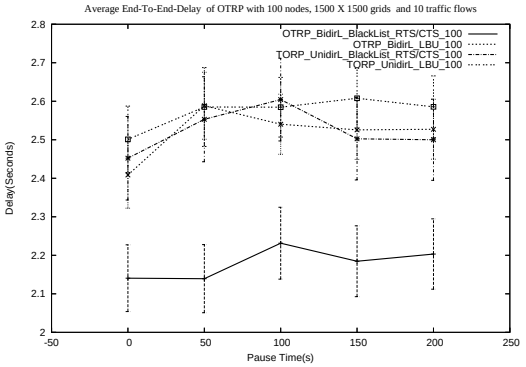


(c) OH

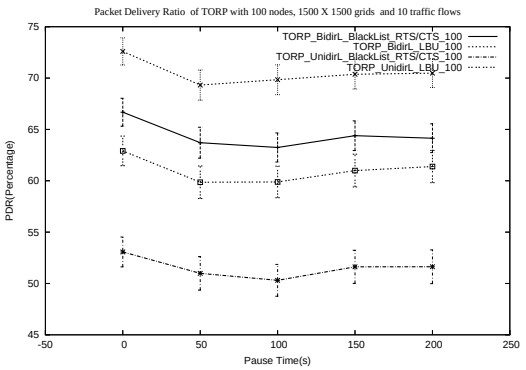


(d) Ret

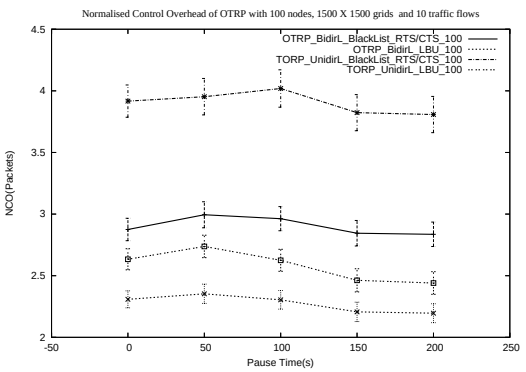
Figure 9: compare LBU to BlackList_CTS/RTS under both bidirectional and unidirectional links with AODV and 100 nodes



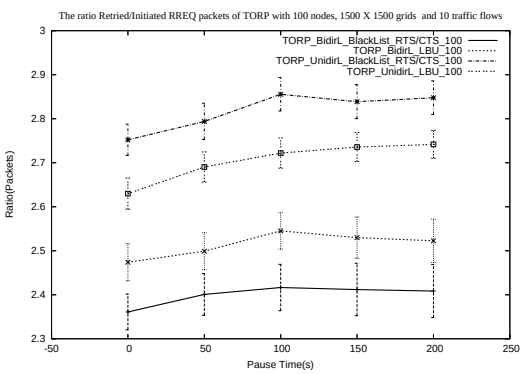
(a) Delay



(b) PDR



(c) OH



(d) Ret

Figure 10: compare LBU to BlackList_CTS/RTS both bidirectional and unidirectional links with OTRP and 100 nodes

work with homogeneous MANET where nodes have similar transmission power and the occurrence of unidirectionality is low. However, LBU outperforms Blacklist_RTS/CTS strategy in term of PDR and NCO under both unidirectional and bidirectional links, see(Figure 9(b-c) and Figure 10(b-c)). This is because LBU strategy supports AODV and OTRP by filtering incoming RREQ packets where not all incoming packets are processed. In other words, incorrect information about first hop neighbours is avoided using LBU. Moreover, our strategy provides sufficient routing information about 2-hop neighbours by solving unidirectional links. In homogeneous MANET where bidirectional links are assumed to exist between any pair of nodes, LBU performs more efficiently than Blacklist_RTS/CTS. AODV_LBU increases PDR by 2% and (see Figure 9(b)) while OTRP_LBU increases PDR by 10% (see Figure 10(b)). Although locations of last 2 hops are attached with RREQ packet in LBU, NCO is improved comparing to Blacklist_RTS/CTS as shown in Figure 9(c) and Figure 10(c) where AODV_LBU and OTRP_LBU reduces NCO by 0.8. However, delay with LBU is higher than Blacklist_RTS/CTS for both protocols where the number of unidirectional link is low under bidirectional links, as shown in (Figure 9(a) and Figure 10(a)) . This is because if unidirectional link is exist between forwarding node and its relay, this will reduce rebroadcasting area, which may increase Ret_Ratio and then consequently increases delay as shown in Figure 9(d) and Figure 10(d). However, detecting unidirectional links and resolving it immediately can guarantee a reliable path to route data, which explains the improvement in PDR and NCO. In heterogeneous MANET, nodes with different transmission are exist. Therefore, high percentage of unidirectional links occur. In both protocols, LBU resolves this problem without any increasing of NCO or delay comparing to Blacklist_RTS/CTS strategy or other strategies as you can see in Figure 9 and Figure 10. This is because LBU detects and immediately resolves any unidirectional links that may occur in the first RREQ_RETRIAL (see Figure 9-(d) and Figure 10-(d)) comparing to Blacklist strategy where unidirectional links are avoided and some nodes are blocked. Therefore, AODV_Blacklist_RTS/CTS and OTRP_Blacklist_RTS/CTS consume nearly 2 and 3 respectively out of 3 RREQ_RETRIALS to find bidirectional paths to route the data. This will increase delay as shown in Figure 10(a). Unlike AODV, the number of rebroadcasting nodes is eliminated in OTRP, which reduce rebroadcasting area and hence OTRP requires more RREQ_RETRIAL. Therefore, generally the delay with OTRP is sightly higher than AODV but OTRP_LBU has constant delay. As RTS/CTS is used too, Blacklist_RTS/CTS consequently increase NCO by 1.5 and decrease PDR by at least 6% as in Figure 9-(b and c) and Figure 10-(b and c) respectively.

VI. Conclusion

In this paper, LBU is proposed to resolve unidirectional link in MANET. Instead of dropping duplicated RREQ packet, each incoming RREQ packet is used to filter routing information of neighbours under unidirectionality. LBU and Blacklist_RTS/CTS are applied on top of AODV and OTRP. LBU outperforms Blacklist with RTS/CTS strategies under homogeneous and heterogeneous MANET in term of PDR and NCO, and without increasing delay.

References

[1] H. A. Amri, M. Abolhasan, and T. Wysocki, "Scalability of manet routing protocols for heterogeneous and homogenous networks," *Computers and Electrical Engineering*, 2009.

[2] M. Zuhairi and D. Harle, "Dynamic reverse route in ad hoc on demand distance vector routing protocol," in *Wireless and Mobile Communications (ICWMC), 2010 6th International Conference on*, pp. 139 –144, sept. 2010.

[3] G. Wang, D. Turgut, L. Bölöni, Y. Ji, and D. C. Marinescu, "A simulation study of a mac layer protocol for wireless networks with asymmetric links," in *Proceedings of the 2006 international conference on Wireless communications and mobile computing*, IWCMC '06, (New York, NY, USA), pp. 929–936, ACM, 2006.

[4] M. Abolhasan, J. Lipman, and J. Chicharo, "A routing strategy for heterogeneous mobile ad hoc networks," vol. 1, (New York, NY 10016-5997, United States), pp. 13 – 16, 2004.

[5] A. Abbas and B. Jain, "Topology control in heterogeneous mobile ad hoc networks," in *Personal Wireless Communications, 2005. ICPWC 2005. 2005 IEEE International Conference on*, pp. 47 – 51, jan. 2005.

[6] T. Maekawa, H. Tada, N. Wakamiya, M. Imase, and M. Murata, "An ant-based routing protocol using unidirectional links for heterogeneous mobile ad-hoc networks," *Wireless and Mobile Communications, 2006. ICWMC '06. International Conference on*, pp. 43–43, July 2006.

[7] W. Liu, Y. Zhang, and Y. Fang, "Conserve energy through multiple-packets transmission in heterogeneous mobile ad hoc networks," vol. 2005, (Piscataway, NJ 08855-1331, United States), pp. 1605906 –, 2005.

[8] S. Das, C. Perkins, and E. Royer, "Ad Hoc On Demand Distance Vector (AODV) Routing," in *Internet Draft, draft-ietf-manet-aodv-11.txt*, (work in progress), 2002.

[9] J.-B. Lee, Y.-B. Ko, and S.-J. Lee, "Euda: detecting and avoiding unidirectional links in ad hoc networks," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 8, pp. 63–67, October 2004.

[10] H. A. Amri, M. Abolhasan, and T. Wysocki, "On optimising route discovery in absence of previous route information in manets," in *To appear in IEEE 69th Vehicular Technology Conference VTC2009-Spring (IEEE VTC)*, 2009.

 ISBN: 978-1-61208-203-5

Different Criteria of Selection for Quantized Feedback of Minimum-Distance Based MIMO Precoder

Ancuta Moldovan*, Ghadir Madi†, Baptiste Vrigneau†, Tudor Palade* and Rodolphe Vauzelle†
*Technical University of Cluj-Napoca, Communications Department
400027 Cluj-Napoca, Romania
Email: firstname.lastname@com.utcluj.ro
†University of Poitiers, XLIM-SIC Department
86962 Futuroscope Cedex, France
Email: firstname.lastname@univ-poitiers.fr

***Abstract*—The achievable performance of a multi-antenna communication system relies on the amount of information describing the channel available at the transmitter and at the receiver. In this paper, we propose different criteria for selecting the optimal precoding matrix from the codebook, in a limited feedback spatial multiplexing system. The distortion function considered to design the codebook aims to maximize the minimum Euclidean distance between signal points at the receiver side. The proposed approaches are compared considering the bit error rate performance, under the constraint of preserving the same rate on the feedback channel. Since the signal adaptation must be performed in real-time, the computational complexity involved in performing the selection is also evaluated for each method. The simulation results show that the proposed solutions are promising methods for quantization, as they ensure very small loss compared to the ideal case.**

***Index Terms*—channel state information; limited feedback; MIMO precoders; precoder selection.**

I. INTRODUCTION

Exploiting propagation diversity, by using multiple antennas at the transmitter and receiver, promises high-capacity and high-quality wireless communication links. In most of the multiple-input multiple-output (MIMO) applications, the channel state information (CSI) at the receiver is needed to perform the equalization and the detection tasks; furthermore, it is known that some degrees of channel knowledge at the transmitter can further boost the system performance through a channel-aware precoding operation of the data streams.

Assuming perfect CSI is available at the transmitter side, different solutions, that optimize pertinent criteria, were proposed in the literature: minimum mean square error (MMSE) [1], signal-to-noise (SNR) maximization at the receiver side ($\max$-SNR) [2], capacity maximization [3] or maximization of the received minimal symbol vector distance ($\max - d_{\min}$) [4]. The limitation of these methods is the fact that obtaining accurate CSI at the transmitter (CSIT) requires to assign a substantial amount of resources on the feedback channel, which may reduce the efficiency of the communication system.

In order to meet the bandwidth requirements on the feedback channel, an efficient quantization of the CSI is mandatory. In [5], Love et al. present a general overview of feedback quantization in various transmitter adaptation schemes. Among these, the quantization of the beamforming vector leads to a significant advancement in feedback techniques. The solution in [6] assumes that the receiver chooses the precoding matrix from a finite cardinality codebook, designed off-line and known at both sides of the wireless communication link. The challenges associated with this quantization scheme are the design of the codebook and the criterion to select the optimal precoding matrix from the codebook. The framework used for the limited feedback beamforming is related to the Grassmann sub-space packing problem, approach that was shown to ensure the outage minimization, the SNR and the rate maximization.

Based on these insights, in [7], we have made the first steps for the design of a new quantization scheme for the $\max - d_{\min}$ precoder. The reason why we have opt for the $\max - d_{\min}$ precoder is the fact that, under the assumption of perfect CSI, it achieves good performance in terms of bit error rate (BER), providing a significant gain of SNR compared to others precoders [8]. For the quantized feedback, the codebook design method and the selection of the optimal precoding matrix are based on the maximization of the minimum Euclidean distance. The proposed method seams to give good results in MIMO systems with two transmit antennas.

In this paper, based on the codebook designed in [7], we propose different functions for the selection of the optimal precoding matrix from the codebook, for a maximum likelihood (ML) detection. We present results obtained by evaluating the BER in a (2,2) MIMO system, while taking into account the feedback rate and the computational complexity of the proposed solutions.

The remainder of the paper is structured as follows. In Section II, the system model and the $\max - d_{\min}$ precoder are presented. Section III introduces the codebook design method. In Section IV, different selection criteria are presented and

 ISBN: 978-1-61208-203-5

evaluated , considering the same rate on the feedback channel. Section V resumes the conclusions and states the future work.

II. System model and the $\max -d_{\min}$ precoder

We consider a MIMO system with n_T transmit and n_R receive antennas operating over an i.i.d. Rayleigh flat-fading channel and b independent data streams are to be transmitted ($b \leq \min(n_T, n_R)$). Assuming perfect knowledge of the channel state information at both sides of the wireless link, a precoder and a decoder matrices, $\mathbf{F}(n_T \times b)$ and $\mathbf{G}(b \times n_R)$, can be designed, so that the basic system model is:

$$\mathbf{y} = \mathbf{GHFs} + \mathbf{Gn} \tag{1}$$

where $\mathbf{H}$ is the $n_R \times n_T$ channel matrix, $\mathbf{s}$ is the $b \times 1$ vector of transmitted symbols and $\mathbf{n}$ is the $n_R \times 1$ additive white Gaussian noise (AWGN) vector.

By using the following decomposition $\mathbf{F} = \mathbf{F}_v\mathbf{F}_d$ and $\mathbf{G} = \mathbf{G}_v\mathbf{G}_d$, the input-output relation (1) can be rewritten as:

$$\mathbf{y} = \mathbf{G}_d\mathbf{H}_v\mathbf{F}_d\mathbf{s} + \mathbf{G}_d\mathbf{n}_v \tag{2}$$

where $\mathbf{H}_v = \mathbf{G}_v\mathbf{HF}_v = \mathrm{diag}(\sigma_1, ..., \sigma_b)$ is the $b \times b$ eigen-mode channel matrix, with σ_i representing each sub-channel gain, in a decreasing order; $\mathbf{n}_v = \mathbf{G}_v\mathbf{n}$ is the $b \times 1$ additive noise vector on the channel eigen-mode; the unitary matrices $\mathbf{F}_v$ and $\mathbf{G}_v$ are chosen so as to diagonalize the channel and to reduce the dimension to b.

As only the ML detection is considered at the reception, the decoder $\mathbf{G}_d$ has no impact on the performance and it is considered to be $\mathbf{G}_d = \mathbf{I}_b$, where $\mathbf{I}_b$ is the $b \times b$ identity matrix. Regarding the transmit precoder, the optimization under the $\max -d_{\min}$ criterion gives the matrix $\mathbf{F}_d$:

$$\mathbf{F}_d = \arg\max_{\mathbf{F}_d} \min_{\mathbf{s}_k, \mathbf{s}_l \in C^b, \mathbf{s}_k \neq \mathbf{s}_l} ||\mathbf{H}_v\mathbf{F}_d(\mathbf{s}_k - \mathbf{s}_l)|| \tag{3}$$

where $\mathbf{s}_l$ and $\mathbf{s}_k$ are 2 symbols vectors whose entries are elements of the received constellation C.

The solution of (3) is difficult since it involves the computation of the minimum distance. A very exploitable solution was given in [4] for two independent data streams, b = 2 and a 4-QAM, with a spectral efficiency of η = 4bits/s/Hz. If we consider the 2-dimensional virtual channel $\mathbf{H}_v = \mathrm{diag}(\sigma_1, \sigma_2)$, it can be totally defined by two parameters: a positive real parameter $\rho = \sqrt{\sigma_1^2 + \sigma_2^2}$ which is the channel gain and $\gamma = \arctan(\sigma_2/\sigma_1)$ the channel angle, $\pi/4 \geq \gamma > 0$.

The precoding solution is SNR-independent and is based on the value of the channel angle:

- if $0 \leq \gamma \leq \gamma_0$

$$F_d^{d_{\min}} = F_{r1} = \sqrt{E_T} \begin{bmatrix} \sqrt{\frac{3+\sqrt{3}}{6}} & \sqrt{\frac{3-\sqrt{3}}{6}} e^{\frac{i\pi}{12}} \\ 0 & 0 \end{bmatrix} \tag{4}$$

- if $\gamma_0 \leq \gamma \leq \pi/4$

$$F_d^{d_{\min}} = F_{octa} = \sqrt{\frac{E_T}{2}} \begin{bmatrix} \cos\psi & 0 \\ 0 & \sin\psi \end{bmatrix} \begin{bmatrix} 1 & e^{\frac{i\pi}{4}} \\ -1 & e^{\frac{i\pi}{4}} \end{bmatrix} \tag{5}$$

where $\psi = \arctan\frac{\sqrt{2}-1}{\cos\gamma}$ is related to the power allocation and $\gamma_0 = \arctan\sqrt{\frac{3\sqrt{3}-2\sqrt{6}+2\sqrt{2}-3}{3\sqrt{3}-2\sqrt{6}+1}} \simeq 17.28^0$ is a constant threshold, computed by considering that the two forms of the precoder provide the same $d_{\min}$.

A power constraint has to be satisfied, the average transmit power being limited to E_T, so $\mathrm{trace}(\mathbf{F}_d\mathbf{F}_d^*) = E_T$.

III. Codebook design based on channel statistics

The approach is related to the Grassmannian line packing technique, presented in [6], for the design of the beamforming codebook $\mathcal{F}_v$=$\{\widetilde{\mathbf{F}}_{v_1}, ..., \widetilde{\mathbf{F}}_{v_N}\}$, where N is the cardinality of the codebook and $n = \log_2$N is the number of feedback bits. The method used for the design of $\mathcal{F}_v$ can be applied only for an uncorrelated Rayleigh fading channel and it is independent on the channel realization and on the number of receive antennas. In order to eliminate these limitations, in [7] we have presented a new limited feedback scheme that involves the quantization of the $\max -d_{\min}$ precoder. The codebook is empirically design from simulation, as presented in what follows.

Design criterion: Using a sufficiently large number of channel realizations, for each element $\widetilde{\mathbf{F}}_{v_i}$ from $\mathcal{F}_v$, $i = 1$:N, we determine $\mathbf{F}_d$ that maximizes $d_{\min}$:

$$\arg\max_{\mathbf{F}_d} \min_{\mathbf{s}_k, \mathbf{s}_l \in C^b, \mathbf{s}_k \neq \mathbf{s}_l} ||\mathbf{G}_v\mathbf{H}\widetilde{\mathbf{F}}_{v_i}\mathbf{F}_d(\mathbf{s}_k - \mathbf{s}_l)|| \tag{6}$$

The product given by each pair $(\widetilde{\mathbf{F}}_{v_i}, \mathbf{F}_d)$ represents an entry in the new codebook $\mathcal{F}$=$\{\widetilde{\mathbf{F}}_1, ..., \widetilde{\mathbf{F}}_N\}$. The elements are dependent on the configuration of the MIMO system thanks to the design of the $\max -d_{\min}$ precoder. For certain channel realizations, meaning values of the channel angle γ above the threshold γ_0, the new entries are also dependent on the matrix describing the MIMO channel. Since the parameter that connects the matrix $\mathbf{H}$ and the quantized values of the precoder is the channel angle, to each element in $\mathcal{F}$ we associate a value of γ related to the $\mathbf{F}_d$ precoding matrix. All the quantized values of γ are contained in a second codebook $\mathcal{A} = \{\widetilde{\gamma}_1, ..., \widetilde{\gamma}_N\}$.

IV. Selection criteria

In this section, we are providing different design approaches for the selection of the optimal precoding matrix from the codebook. Numerical results are presented assuming a (2,2) MIMO system, over which two independent data streams are transmitted and a 4-QAM modulation.

 ISBN: 978-1-61208-203-5

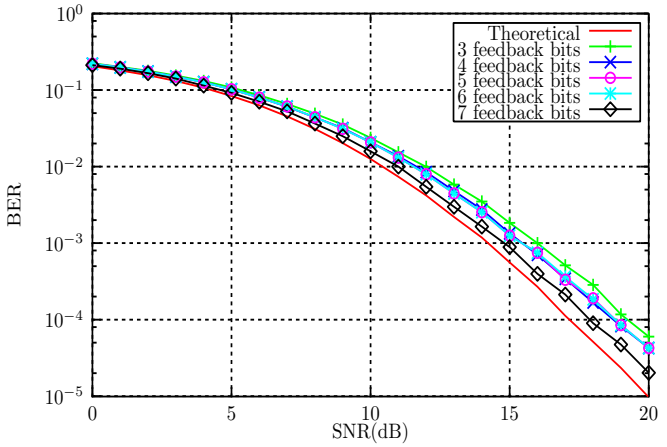


Figure 1: Performance of the $d_{\min}$ criterion dependent on the number of feedback bits

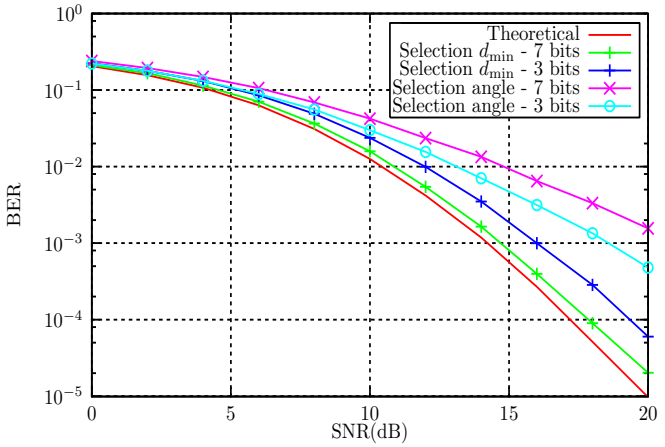


Figure 2: Performance of the $d_{\min}$ and the angle selection criteria dependent on the number of feedback bits

A. $d_{\min}$ selection

The selection criterion maximizes the minimum Euclidean distance between signal points on the received constellation. For the given codebook $\mathcal{F}$, the receiver encodes as follows:

$$\arg\max_{\widetilde{\mathbf{F}}} \min_{\mathbf{s}_k,\mathbf{s}_l \in C^b, \mathbf{s}_k \neq \mathbf{s}_l} ||\mathbf{G}_v \mathbf{H} \widetilde{\mathbf{F}}_i (\mathbf{s}_k - \mathbf{s}_l)|| \quad (7)$$

In Figure 1 are depicted the BER results for the proposed limited feedback scheme. The results for the perfect feedback are also plotted for comparison. For a BER of 10^{-4} and considering 7 feedback bits, the loss relatively to the ideal case is 0.7 dB. Reducing the feedback rate at 3 bits leads to a significant degradation of the BER performance, of about 2 dB, which is considered to be unacceptable.

The selection provides good performance with a sufficient number of feedback bits, but the computation of $d_{\min}$ depends on both the modulation's order and on the channel's statistics. Moreover, in equation (7) it is necessary to consider all possible error vectors in searching for the optimal precoding matrix. In [4] it was shown that for a 4-QAM modulation there is a number of 14 difference vectors that must be considered, but it increases significantly with the modulation's order. Since the selection of the precoder must be performed in real-time, it is necessary to reduce the computational complexity. In what follows, we will consider other selection functions mainly based on the parameters involved in the singular value decomposition (SVD) of the channel matrix.

B. Angle selection

The channel angle is a parameter that can be used to relate the current channel realization with the codebook entries. The new criterion is based on the channel angle and it is intended to minimize the difference between the actual channel angle γ, and the quantized values $\widetilde{\gamma}_i$, from the dictionary $\mathcal{A}$. The function returns the index k of the selected codebook entry, that will be used to determine the precoding matrix $\widetilde{\mathbf{F}}_k$ from the associated codebook $\mathcal{F}$.

$$k = \min_{i=1:N} |\gamma - \widetilde{\gamma}_i| \quad (8)$$

One advantage of the criterion is the fact that it avoids the computation of the minimum distance, but it requires the diagonalization of the channel matrix. Since traditional SVD algorithms involve costly arithmetic operations [9], increased efficiency can be obtained by making use of hardware oriented arithmetic techniques, based on the CORDIC (COrdinate Rotation Digital Computer) arithmetic [10].

The simulation results depicted in Figure 2 show a significant performance degradation, in terms of BER, when the angle selection is applied. Based on the codebook construction, we assumed that the BER depreciation is caused by values of $\gamma \leq \gamma_0$. The next step was to determine to what extent these values influence the BER. So, using a set of 10^5 samples of the channel matrix, we estimated the cumulative distribution function (cdf) for a (2,2) MIMO system. From the results in Figure 3, the probability that the channel angle is bellow the threshold is $P(\gamma \leq \gamma_0) \simeq 0.44$ This means that, in almost half of the cases so there is no connection between the codebook entries and the actual channel realization, so for this region, the quantization scheme is inefficient.Moreover, in this case, the higher the number of codebook entries, the lower is the BER performance since the quantized values of the precoder are independent on the channel matrix.

In order to highlight the influence of the $\max - d_{\min}$ precoder forms on the system's performance, we partition the channel space based on the values of the channel angle. In Figure 4, the BER plots are computed separately, considering either channel matrices with $\gamma \leq \gamma_0$, or with $\gamma \geq \gamma_0$ and a feedback rate of 3 bits. The optimal BER values, for the same channel statistics, are also plotted for comparison. The

 ISBN: 978-1-61208-203-5

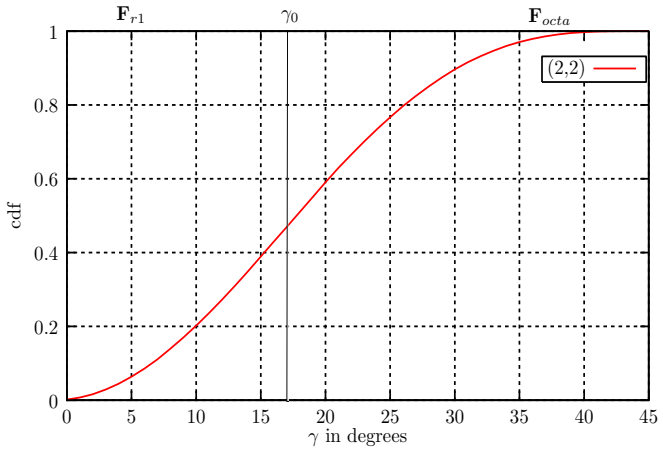


Figure 3: Cumulative distribution function for γ in a (2,2) MIMO system

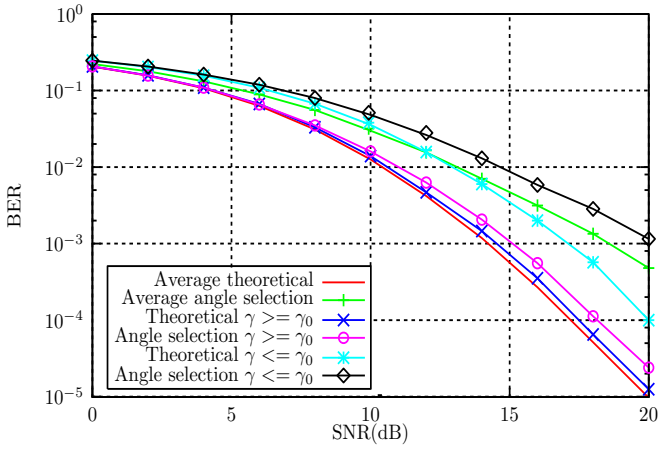


Figure 4: Performance of the angle selection criterion based on the value of the channel angle - 3 feedback bits

results lend credence to our observation that the angle based selection is to be used only for channel matrices with γ above the threshold.

C. Hybrid selection

Based on the results obtained with the previous two selection criteria, we propose a new function that combines them, depending on the value of the channel angle. So, for values of the channel angle bellow the threshold the $d_{\min}$ criterion is used to select the optimum precoder from the codebook, while for values of the channel angle above the threshold, the selection is based on γ.

The results obtained with the hybrid selection are depicted in Figure 5. It can be observed that, with a 3-bits feedback rate, the performance is comparable with the one obtained with the $d_{\min}$ criterion and the same number of bits on the feedback channel. In what concerns the complexity, the channel diagonalization is needed, but for more than half of the scenarios the $d_{\min}$ computation is avoided.

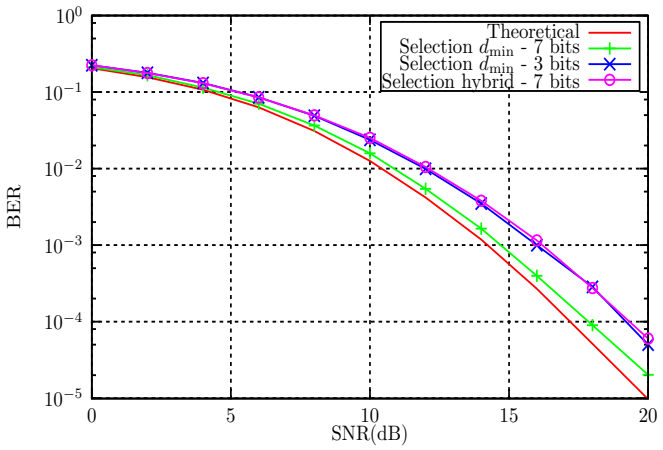


Figure 5: Performance of the hybrid and the $d_{\min}$ selection criteria

D. $\max -\sigma_1$ selection

In the hybrid selection, we focus on channel matrices with the angle above the threshold. In what follows, we propose a selection criteria that takes into account also the channel matrices corresponding to the $\mathbf{F}_{r1}$ form of the $\max -d_{\min}$ precoder. Since the $P(\gamma \leq \gamma_0) \simeq 0.44$ it means that in almost half of the channel realization the precoder uses only the first virtual sub-channel to transmit the data symbols. Based on this observation, we propose a quantization scheme that emphasizes the use of the first singular value.

The same approach as in Section III is considered for the codebook generation, but the criterion is intended to maximize the first singular value $\max -\sigma_1$ from the virtual channel matrix given by:

$$\sigma_1(\arg\max_{\mathbf{F}_d}(\mathbf{G}_v\mathbf{H}\widetilde{\mathbf{F}}_{v_i}\mathbf{F}_d)) \quad (9)$$

Regarding the selection of the optimal precoder from the codebook, the same criterion that maximizes the first singular value σ_1 is applied. Compared to the $d_{\min}$ criterion, the new function does not require considering the difference vectors set involved in the computation of the minimum distance. The BER results are depicted in Figure 6. For a BER of 10^{-4}, with a 7 bits feedback rate, the loss relative to the $d_{\min}$ based quantization scheme is around 0.6 dB. The rate reduction at 3 feedback bits adds no extra SNR loss.

In the attempt to reduce the complexity due to the computation of the minimum distance, while keeping a low rate on the feedback channel, we have introduced different alternative methods for selecting the optimal precoder from the codebook. The new functions are related to the design of the $\max -d_{\min}$ precoder and are dependent on the singular value decomposition of the channel matrix. For channel realizations characterized by a value of the channel angle higher than the threshold we have proposed a solution based on γ. Another

 ISBN: 978-1-61208-203-5

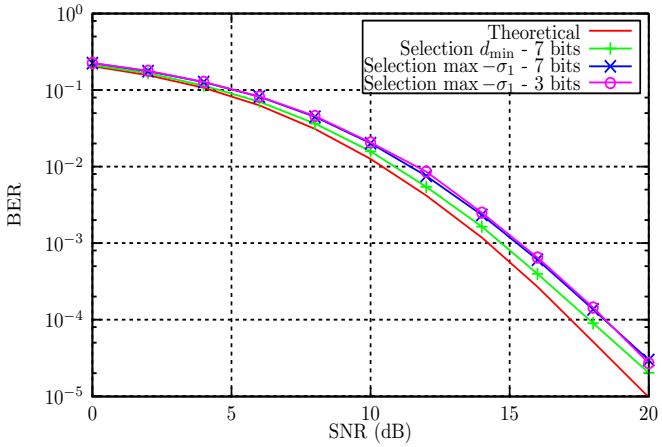


Figure 6: Performance of the $d_{\min}$ and the $\max(\sigma_1)$ selection criteria

solution relies on emphasising the use of the first sub-channel from the virtual channel matrix to transmit the data.

The proposed distortion functions ensure wireless connections with different performance in terms of reliability, depending also on the number of feedback bits. When choosing a certain selection method one must take into account the complexity of the involved computations. It must be stated that for larger MIMO systems is even more important to consider the complexity.

V. Conclusion

In this paper, we propose a quantization scheme for uncorrelated Rayleigh fading channels, applied to the $\max -d_{\min}$ precoder. The codebook design method and the selection of the optimal precoder from the codebook are based on the maximization of the minimum Euclidean distance on the received constellation. If the two forms of the $\max -d_{\min}$ precoder are considered in the development of the quantization scheme, the simulation results have shown that, with a reduced feedback rate, 3 bits, the method ensures a very small loss compared to the ideal case.

References

[1] A. Scaglione, S. Barbarossa, G. Giannakis and H. Sampath, "Optimal designs for space-time linear precoders and decoders",*IEEE Transactions on Signal Processing* vol. 50, no. 5, pp. 1051-1064, May 2002.

[2] P. Stoica and G. Ganesan, "Maximum-SNR spatial-temporal formatting design for MIMO channels",*IEEE Transactions on Signal Processing* vol. 50, no. 12, pp. 3036-3042, 2002.

[3] I.E. Telatar", "Capacity of multi-antenna Gaussian channels",*European Transactions on Telecommunications* vol. 10, pp. 585-595, 1999.

[4] L. Collin, O. Berder, P. Rostaing and G. Burel, "Optimal minimum distance based precoder for MIMO spatial multiplexing systems",*IEEE Transactions on Signal Processing* vol. 52, no. 3, pp. 617-627, 2004.

[5] D.J. Love, R.W. Hearth, V.K.N. Lau, D. Gesbert, B.D. Rao and M. Andrews, "An overview of limited feedback in wireless communications systems",*IEEE Journal on Selected Areas in Communications* vol. 28, no. 8, pp. 1341 - 1365, October 2008.

[6] D.J. Love and R.W. Hearth, "Limited feedback unitary precoding for spatial multiplexing systems",*IEEE Transactions on Signal Processing* vol. 51, no. 8, pp. 2967-2976, August 2005.

[7] A. Moldovan, G. Madi, B. Vrigneau, T. Palade and R. Vauzelle, "SVD algorithms and quantization applied to MIMO $\max -d_{\min}$ based precoder",*Proceedings of EURASIP Workshop on Signal Processing and Applied Mathematics for Electronics and Communications, SPAMEC 2011*, pp. 5-8, 2011.

[8] B. Vrigneau, J. Letessier, P. Rostaing, L. Collin and G. Burrel, "Statistical comparison between $\max -d_{\min}$, max-SNR and MMSE precoders", *40th Asilomar Conference on Signals, Systems and Computers*, Oct. 2006.

[9] J.R. Cavallaro and F.T. Luk, "CORDIC Arithmetic for an SVD Processor," *in Proc. 8th Symp Computation Arithmetic*, pp. 113-120, May 1987.

[10] J.E. Volder, "The CORDIC trigonometric computing technique," *IRE Transactions on Electronic Computers* vol. 8, no. 3 pp. 330-334, Sep. 1959.

 ISBN: 978-1-61208-203-5

On Browsing Behavior-based Traffic Model of Mobile Internet

Hong Tang, Xiang-yue Kong, Lu Wang, Yu Wu
Network and Computation Research Center.
Chongqing University of Posts and Telecommunications.
Chongqing, China.
emails{tanghong@cqupt.edu.cn, kongxiangyue@gmail.com, lue_wang@qq.com, wuyu@cqupt.edu.cn}

***Abstract*—With the rapid rising of the number of cellular phone users, accessing the Internet via handset devices is becoming a standard configuration in terms of network activities and people's daily life, resulting in an ever-increasing usage of the Mobile Internet. Yet, there were little knowing about how users' behavior on Mobile Internet is. Considering most Mobile Internet users satisfy browsing, the browsing traffic is focused in this paper. Through studying on users' browsing behavior on Mobile Internet with an extended On-Off model to understand generation mechanism of browsing traffic, this paper proposed a browsing traffic model in which such results as self-similarity of traffic volume and following Pareto and Weibull distribution of File Size, View Time and WAP Gateway Response Time were found out by investigating real data sets. This paper launches a primary research on traffic model for Mobile Internet browsing behavior.**

Keywords-Mobile Internet;User behavior;Browsing Traffic Model;K-S test

I. INTRODUCTION

Mobile Internet has become a profitable and promising business. According to CNNIC's (China Network Information Center) report [1], there were about 35,558 netizens investigated had Mobile Internet surfing experience with mobile phone in 2011, while 11,760 in 2008 which have increased about 3 times in these years. The report also shows that 62.1% and 60.9% Mobile Internet netizens habitually use news and search service [1], respectively, which means that browsing traffic is the main stream of Mobile Internet traffic now in China.

However, there were few studies on user behaviors on Mobile Internet while such throne market. Some previous researches [2] [3] [4] [5] focused on WAP-based (Wireless Application Protocol) mobile network behaviors. The model of the WAP traffic generated by requesting web pages that reply with Wireless Makeup Language (WML) files in General Packet Radio Service (GPRS) network was studied. Varga et al. [2] provided a traffic model based on long-term, live measurements, and observations to estimate the user behavior and the workload in GPRS network. Irene C. Y. Ma et al. [3] constructed a model of WAP traffic based on a number of user scenarios to study the characteristics of the WAP traffic. Toshihiko Yamakami [4,5] studied user behaviors on mobile Internet in Japan. By examining the long-term mobile Internet user transaction logs, he analyzed the long-term usage pattern to study the notion of user "age" (the length of user experience [4]) and explored regularity measures to track user behaviors based on an ad-hoc assumption that the user loyalty relates to the web visit regularity [5]. In some newest research on Mobile Internet traffic, Lymberopoulos et al. [20] proposed to use a machine learning approach based on stochastic gradient boosting techniques to efficiently model the signature of Mobile Internet users whose web access traces were analyzed. Chuan Xu et al. proposed a new method of measuring the similarity of daily clicks distribution by Pearson Correlated Coefficient and introduced various means to describe the heterogeneity of clicks distribution both in users and in websites [21]. Some interesting results (like users' obeying 20/80 rule) were found in [20][21].

Being a basic theoretical issue, the research of traffic model is valuable for WAP/Web site owners and network operators and also useful in study on future network. Most users of Mobile Internet, as page browsers, are satisfied with viewing page such as news and search. Thus, browsing traffic (or HTTP traffic) is the dominating traffic of Mobile Internet. This paper aims to build a traffic model of page browsing for describing the browsing traffic behaviors on Mobile Internet. The contributions of this paper are summarized as follows:

- An extended On-Off model is used to help understanding the mechanism of traffic generation of page browsing behavior on Mobile Internet. Traffic Volume, File Size, View Time and WAP Gateway Response Time were determined to describe the traffic model of page browsing behaviors.
- We analyzed the parameters of model through a real data set from methods such as Hurst coefficient [12] and K-S test [2].

This paper is organized as follow: We constructed a browsing traffic model in Section II. In Section III, data sets obtained in two years is presented. With these data sets, we worked out the analysis of the browsing traffic model by Hurst coefficient and K-S test with data we contained in Section IV. Finally, this paper is concluded in Section V.

II. MODEL CONSTRUCTION

A. Communication Mode

In user's browsing behavior on Mobile Internet, there are three kinds of communication entries involved. They are user, gateway and WAP/Web server as shown in Figure 1. On the Mobile Internet, the requests of users have to be transpond by the gateway before arriving servers. The same thing

 ISBN: 978-1-61208-203-5

happens to the responses from server. Users' browsing pages of WAP site or Web site cause their sending two kinds of requests to servers: WAP browsing refers to the browsing demands for the hypertext pages that respond with the WML language following the WAP suite. Web browsing refers to the browsing demands for the PC-based pages to Web servers but through mobile terminals. Yet either the requests of WAP browsing or Web browsing should get through gateway on Mobile Internet. Thus, the traffic of Mobile Internet has to go through the gateway [2].

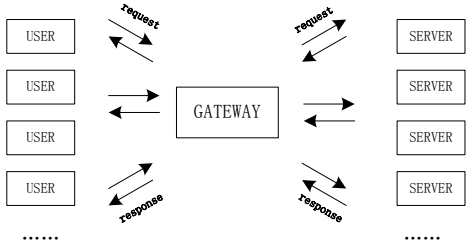


Fig. 1. Communication Mode.

Being in such a special network structure, the traffic behaviors on Mobile Internet are different from that of PC-based Internet.

B. Extended On-Off Model

The On-Off model is introduced in many researches of network traffic [6] [7] [19]. In analyzing the user traffic behaviors, each user is recognized as an On-Off source in which On-State represents traffic generating process and Off-State represents silent period.

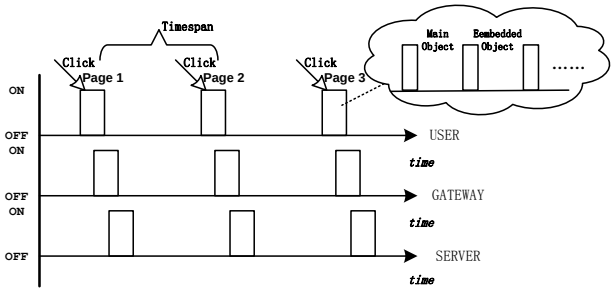


Fig. 2. The Extended On-Off Model.

To explain how the page browsing traffic is generated on Mobile Internet, we extend the On-Off model by adding details from the WAP gateway and WAP/Web server as shown in Figure 2. When a user launches a session, the WAP gateway and WAP/Web server should be ready for serving. It is explicit that user's behavior still can be recognized as two states, On-State for requesting hypertext page started by clicking which generates traffic and Off-State for reading. Consequently, WAP gateway and WAP/Web server would be on On-State for handling requests and sending responses back and Off-State for waiting for serving.

C. Browsing Traffic Model

As stated in previous section, the mechanism of traffic generation of page browsing for the traffic situation depends on the way users switch between On-State and Off-State and network structure. To study browsing traffic we need to answer such questions like how often the user clicks pages, how the pages traffic volume is and how the delay of Mobile Internet network is. Thus we consider 4 factors influencing page browsing traffic which are

- Traffic Volume (TV);
- File Size (FS), including main object size (MS) and embedded object size (ES);
- Viewing Time (VT);
- WAP Gateway Response Time (GRT).

Traffic Volume can directly descript the traffic situation. File Size indicates the traffic volume generated by browsing the traffic of which is also the main traffic of the network. Viewing Time denotes the user reading and clicking frequency. Moreover, WAP Gateway Response Time show how the WAP gateway affects browsing time consumption which is the very procedure delaying the traffic transmission.

With these 4 factors, we primarily establish the browsing traffic model on Mobile Internet. In the rest of this paper, we investigated this model by real data sets.

III. DATA SETS

The dataset in this paper is obtained from the log files of a WAP gateway which belongs to China Telecom. These WAP gateway logs record all the information of Mobile Internet users' online activities for one week from Apr. 5, 2010 to Apr. 11, 2010, and the other week from Apr. 4, 2011 to Apr. 10, 2011. Each record in the log files contains request and response information including Time, Destination Domain, URL, Client IP Address, User Agent, etc.

A basic statistics of these data sets is shown in table, as appendix. The change proportion of WAP traffic and Web traffic can roughly label the change of mobile terminal performance for users tend to abandon WAP page with high-performance terminal which also reflects the change of user behavior pattern.

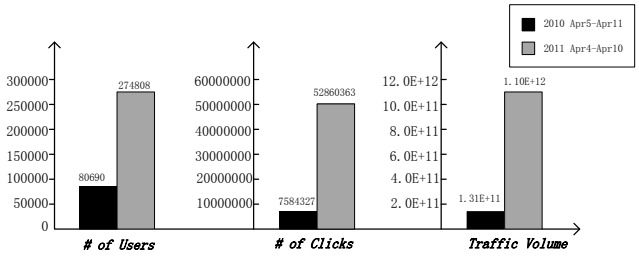


Fig. 3. Data Items Comparison

Here, base on the data obtained, we would like to study data characteristics by statistically analyzing. As shown in Figure 3, we worked out the total number of users, clicks and total flow traffic of the data of the two weeks which have increased about 4 to 8 times in these two years. We can infer that the WAP gateway burdens a higher workload at 2011

comparing 2010, indicating the exploding growing of Mobile Internet.

With the overview of our data sets shown above, we can get an outline about the development of Mobile Internet. However further study is still necessary for practical purposes.

IV. ANALYSIS OF MODEL

A. Traffic Volume

Traffic Volume (TV) is an indicator to simply describe traffic situation of network. Self-similarity is an important characteristic of Traffic Volume as mentioned in [7]. Thus at the beginning, we analyze the elementary daily Mobile Internet traffic pattern to find the law of user behavior by investigating self-similarity of Traffic Volume. However, the traffic analyzed is from macroscopic view, the analysis bases on statistic.

Here are the traffic distributions on statistic for a week in Figure 4. The daily Mobile Internet service condition could easily show the daily user behavior as users browse in day-time and sleep at night. The traffic is high in the daytime because people are active while the traffic stays low after about two hours in the early morning because most of people fall asleep. The relations between traffic of different days are not clear. Yet, we can think over the self-similarity of the Traffic Volume.

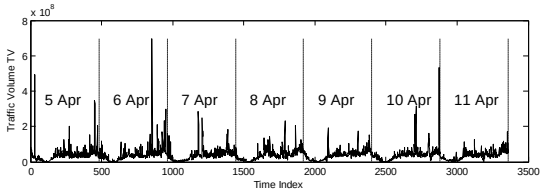


Fig. 4. Traffic Distributions for a Week.

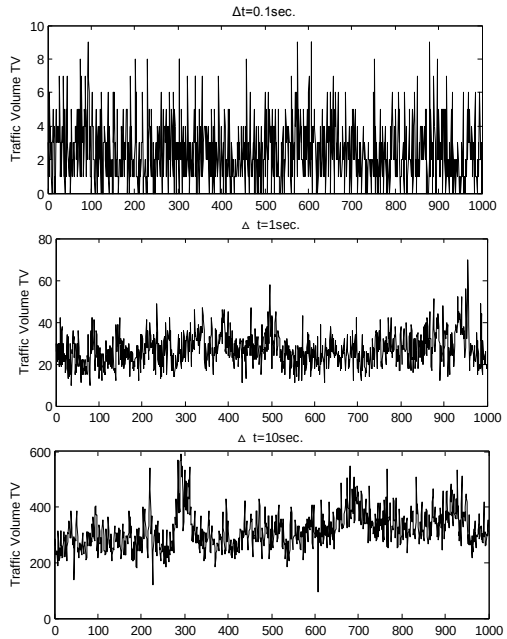


Fig. 5. Sample Aggregation Plot for Mobile Internet Traffic

The self-similarity of network traffic has been proven in many studies [8][9]. The studies on WAP traffic, for the most part, base on simulated WAP traffic or small data set of real trace which only comprise of pure WAP traffic [10][11]. The paper give a self-similarity study based on the data from a large scale data set of real trace of mobile Internet, which not only contains the WAP traffic but also the traffic generated by the connecting between the mobile terminals and Internet.

Mobile Internet traffic was aggregated into various time frames to roughly be observed self-similarity. For 27 consecutive hours of monitored mobile Internet from April 2010 trace, Figure 5 shows a sequence of simple plots of requests which counts for 3 different time frames, each subsequent plot is included in the previous one by increasing the time resolution by a factor of 10 on a random subinterval.

Intuitively, the curves of days in Figure 5 are 'similar' to one anther. To prove the self-similarity of daily mobile Internet traffic, R/S algorithm is used to calculate the Hurst coefficient of daily traffic.

Let $X = \{X_1, X_2, X_3, ..., X_L\}$ be a time series with length L, where the length of each series =10 sec. in the calculation. X is partitioned to be d subsequences with length n obviously, if the value n is definite, there would be $L = d \times n$. The R/S algorithm computes the Hurst coefficient according to [12].

Table I. Hurst Coefficient Distribution for the Week of 2010.

Date	*Num. of time series*	*Scale of partition*	*Num. of fit points*	*Hurst*
5 Apr.	7801	2~3900	2603	0.7984
6 Apr.	7788	2~3894	2598	0.8063
7 Apr.	7805	3~3902	1954	0.7968
8 Apr.	7797	3~3898	1952	0.8114
9 Apr.	7804	3~3902	1954	0.8200
10 Apr.	7748	2~3874	2585	0.7992
11 Apr.	7759	2~3879	2589	0.8038
5-11 Apr	54498	3~27249	13631	0.8169

Hurst found that many time series could be well represented by the relation $(R/S)_n \sim cn^H$, taking the logarithm of both side: $\log(R/S)_n = \log c + H \log n$, where c is a constant. The data fitting method is used after plotting $\log(R/S)_n$ versus $\log n$. The degree of self-similarity is given by H, which is the slope of line of fit above.

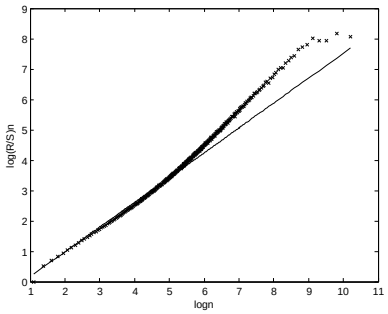


Fig. 6.The Result of R/S for a Week

In this experiment, Mobile Internet traffic was collected consecutively for a week, and the time series was partitioned according to 10 sec. We separately give the results under two different time scales, days and weeks, shown in the table I and the result of R/S for a week is shown in Figure 6.

The parameters of Hurst in Table I are all around 0.8, where a value above 0.5 indicates self-similarity. This compares to the Hurst parameters in the range between 0.76 and 0.83 exhibited by the actual traces captured from web browsing activity [13]. We believe that Mobile Internet in China now being in the preliminary stage, its properties on traffic and some other aspects may be similar with that of PC-based Internet traffic in the early stage.

B. File Size

Though the study on TV above provides the outline of the traffic model, we still need some more details about it. Through the On-Off model, we can understand how a page is generated. Main object, the response to user's clicking, constitutes the page associating with embedded object. As embedded object is requested by browser acting differently from user's clicking, it is necessary to study their traffic pattern respectively. File Size which refers to the size of page objects (such as html, flv and gif) from server is the source of browsing traffic. The investigation to MS and ES can clarify the detail of page browsing behavior.

We can easily obtain File Size by extracting information of DOWNLINK_CONTENT_LENGTH domain of each record in the data set. We distinguish main object record with embedded object record by the analysis of URL domain and Content_Type domain which are recording the URL (Uniform Resource Locator) and type of resource (text, img et) of the object. The two kinds of objects and the results are shown in the Figure 7.

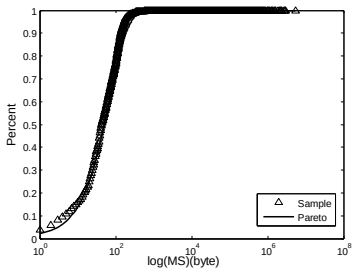


Fig. 7 (a) CDF of MS.

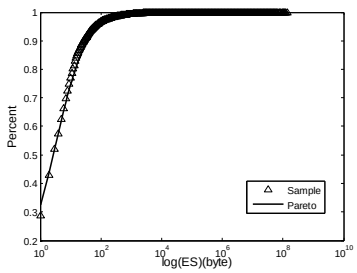


Fig. 7 (b). CDF of ES

Fig. 7. CDF of FS.

Having obtained the CDFs (Cumulative Distribution Function) of File Size, we perform the K-S test on them whose results are also shown in Figure 7 and Table II. In our testing, the samples of our data set follow the Pareto distribution. In Figure 7, we can find that the curves of empirical data fit theoretical curves perfectly. The result comes as the same as [14]. In Table II, the parameters of the fitting Pareto curves are shown.

Table II. K-S test to FS

	Parameter of Fitted Distr.	***Dn***				
	Pareto	***Pareto***	***Norm.***	***Exp.***	***Weibull***	***Loglgst.***
MS	k=0.13 λ =70	0.0353	0.4487	0.1813	0.5524	1
ES	k=1.21 λ =3	0.0747	0.6427	0.1534	0.8663	1

In Figure 7(a), the curve rises directly at about [80,120] which means most MS ranges between 80 to 120 byte. As the main object often carries the information the users going to read, we deem the trend of MS reflects the fact that one page usually provides few information for user on Mobile Internet. This is a normal phenomenon for the small screen of the terminal and specially arrangement of the WAP/Web site owners. The same trend also happens in Figure 7(b) plotting the curve of ES CDF. The ES concentrate on the range between 50 bytes to 200 bytes. It actually can be easily inferred by considering the simplicity of browser and the low capability of terminal. Thus, we believe the Pareto distribution is appropriate for describing the File Size.

C. Viewing Time

The Viewing Time (VT) denotes the timespan that a user browses a web page. The exact Viewing Time actually is difficult to count because we cannot trace the time that the user finishes receiving the web page, so the error is transmission time. However, the transmission time exists in every user click and is transitory compared to the whole VT. We roughly consider the timespan between two user requests in a session to be the view time on mobile Internet. Now, the problem is how to determine each user session (it refers to the course from spanning users' opening to closing browser software.) because the timespans between each user session are contained in the timespans between two user requests.

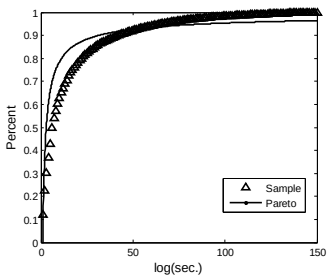


Fig. 8. CDF of VT.

The user session could be determined exactly by Client IP which is randomly assigned. We found Mobile Internet

system randomly assigned Client IP when user logged on Mobile Internet, thus through tracing user log we could easily abandon the timespans between user sessions. However, when the user left idle for a long time (idle time) for some reasons, this timespan was still considered as VT. In PC-based Internet, 79 percent of our test users always scanned any new page they came across; only 16 percent read word-by-word [15], and our trace showed that 97.9% users' VT last within 150 seconds. Thus, VT on Mobile Internet could not be long. So, we give a threshold 150 seconds to determine the idle time, which means that VT lasts within 150 seconds while idle time last longer than it. According to the statistic, VT less than 10 seconds account for 62.85% of all VT of users; viewing time between 10 and 20 seconds is responsible for 17.97%. The CDF of VT is plotted in Figure 8.

In [16] and [17], VT follows Pareto distribution. The sample from our data set, through K-S test mentioned above, is proved to have no best fitted distribution, but the largest vertical distance of Pareto distribution in Figure 8 indeed turn out to be the minimum. It is obviously in Figure 8 that the curve of our sample rises relatively slower than the Pareto distribution. On the average Web page, users have time to read at most 28% of the words during an average visit; 20% is more likely [18]. Comparing to PC-based Internet, the reading habit of user on Mobile Internet is similar. However, because the size of screen for each mobile terminal is much smaller than ordinary computer screen, information delivered from mobile terminals is so limited that users are more willing to read word by word, and taking transmission time into consideration, the VT of mobile Internet users is relative longer than PC-based Internet, which is the reason why the curve of our sample rises slower than Pareto distribution. However, the details of how mobile Internet user read is still not clear, which will be our next work. We believe there is a certain reading habit rule for each user.

D. Gateway Response Time

Because of particular communication structure, the total response time of Mobile Internet is relatively longer than PC-based Internet for one request process. Firstly, the WAP Gateway Response Time (GRT) accounts for a significant part while there is no such period in PC-based Internet. Secondly, computing the Mobile Internet response time for one request need take many environmental factors such as weather conditions, geographic location into consideration. In this part, we observed the response time of the two main elements on Mobile Internet and explored their relations. At the same time, the Viewing Time from Mobile Internet will be made comparison with it on the PC-based Internet.

We observed the GRT for consecutive 168 hours and were looking forward to find its evolution with time. For better observation, the GRT for 168 hours was analyzed by dividing all the response time into several time ranges, and we explored how the different requests distribute in various time ranges, which is shown in Table III.

After exploring the response time in detail, we give a macroscopical description on GRT distribution shown in Figure 9. It could be predicted from Table III, which exhibits most of requests were processed in a certain time range.

Table III. GRT to Different Objects

Time ranges (millisec.)	***Num. of main object***	***Num. of embedded object***	***percentage***
Above 5000	4624	134	2.9%
500~5000	208	38	18.3%
300~500	249	14	5.6%
100~300	7606	226	3.0%
50~100	8488	140	1.6%
30~50	39694	100	0.3%
20~30	101889	318	0.3%
10~20	969056	52111	5.4%
5~10	6290582	2517246	40.0%
0~5	9894219	5515933	55.7%

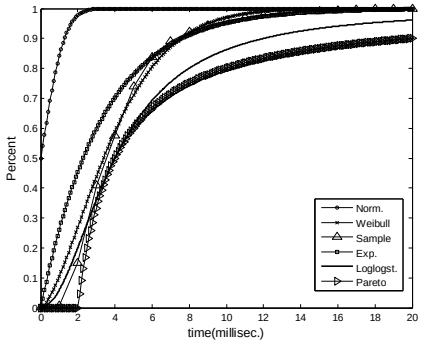


Fig. 10. K-S test to GRT.

Table IV. K-S test to GRT

	Mean	***Std. Dev.***	***Fitted Distr.***	***Parameter***	***α***
Response time	6.343	95.241	\	\	0.05
Normal response time	4.672	2.675	Weibull Distr.	λ =3 k=2	0.05
	Dn				
	Exp Distr.	***Norm. Distr.***	***Weibull Distr.***	***Loglgst Distr.***	***Pareto Distr.***
Response time	8.139	149335	7.003	26.587	3760
Normal response time	1.387	3.849	1.242	4.611	643

In Table IV, the K-S test was applied to measure the data set of the response time and no distribution was shown to be well fitted. The reason, through analysis, is that most of the requests were processed within the "normal" response time while few of requests for some reasons exceeded. Therefore, it is obvious to notice from Figure 10 that the curve is steep at beginning. The normal response time mentioned above refers to the response time within 20 milliseconds. We found that the normal response time well fitted the Weibull distribution through K-S test with $c(\alpha)$=1.358 and significance level of α=0.05, which is shown in table 5 and Figure 10 demonstrate the CDF of aggregated normal response time. By visual inspection, we can find that the

 ISBN: 978-1-61208-203-5

CDF curve of Weibull distribution most closes to that of the sample.

In [2], non-rejects of K-S test occur rarely because the large size of data sets generated by some underlying mechanisms cannot be easily modeled. In our experiment, frankly speaking we didn't make clear of detailed mechanism in WAP gateway but depicted the data set from statistic. We found that the goodness of fitness for a certain function distribution changed with its parameters when applying the K-S test to the data set. So the fitted distribution we give maybe not the best one, but we want to do is to give an appropriate and exact description to our samples and elaborate its properties.

V. CONCLUSION

Considering the different network structures between Mobile Internet and PC-based Internet, we utilized the On-Off model extended to understand how the user's behavior of browsing page influents traffic situation of Mobile Internet. Based on that, we built up a model for the user's page browsing behaviors on Mobile Internet.

We performed an analysis of the traffic model by methods such as Hurst coefficient and K-S test with the data sets collected at the same period in 2010 and 2011 which also indicate a development of Mobile Internet in these two years. It was found that Traffic Volume was with the property of self-similarity; File Size followed Pareto distribution; Gateway Response Time followed Weibull distribution; yet the property of Viewing Time was hard to determine. These results are significantly valuable in assisting network operators to further optimize Mobile Internet network settings.

However the further works is still necessary. In the future, we will study the browsing traffic model established in the paper more deeply. We could try to find out the relation between the parameters of this model. This will help us get a better understanding of traffic situation of Mobile Internet.

ACKNOWLEDGMENT

This work is partially supported by the Natural Science Foundation of the City of Chongqing (grant: CSTC.2009BA2089), Important National Science & Technology Specific Projects-New Generation Broadband Wireless Mobile Communication Network (grant: 2012ZX03006001-004) and Natural Science Foundation of Chongqing Education Commission (grant: KJ100508 and KJ100515).

REFERENCES

[1] China Network Information Center, 29th Statistical Report on Internet Development in China, http:// tech.163.com/ special/ cnnic29/, 2012.1.

[2] T. Varga , Haverkamp B. and Sanders B., Analysis and Modeling of WAP Traffic in GPRS Networks, In ITC Specialist Seminar 2004.

[3] I. C. Y. Ma and J. Irvine, Characteristics of WAP traffic, Wireless Networks, Vol.10, No.1, pp.71-81.

[4] Toshihiko Yamakami, Toward Understanding Mobile Internet User Behavior: A Methodology for User Clustering with Aging Analysis, Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies, 2003.

[5] Toshihiko Yamakami, Classification of Mobile Internet User Behavior Using Qualitative Transition Patterns, Fourth International Conference on Information Technology, 2007. GPGS Networks

[6] Murad S. Taqqu, W. Willinger and R. Sherman, Proof of a Fundamental Result in Self-Similar Traffic Modeling, Computer Communication Review – CCR, 1997, Vol. 27, No. 2, pp.5-23.

[7] J. Lee and M. Gupta, A new trafficmodel for current user web browsing behavior, Tech. Rep., Intel, Santa Clara, Calif, USA, 2007.

[8] Will E. Leland, Murad S. Taqqu, Walter Willinger and Daniel V. Wilson, On the Self-Similar Nature of Ethernet Traffic, IEEE/ACM TON, Feb. 1994 Vol.2, No.1, pp.1-15.

[9] M. E. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic: evidence and possible causes, ACM SIGMETRICS Performance Evaluation Review, May 1996, Vol.24, No.1, pp.160-169.

[10] Wang Yu, Zhu Chun-mei and Wu Wei-ling, Traffic model of WAP over GPRS, Communication Technology Proceedings, 2003, pp 1674-1677.

[11] Irene C. Y. Ma and James Irvine, Characteristics of WAP traffic, Wireless Networks,Vol.10 No.1, pp.71-81.

[12] Fu Leiyang, Wang Ruchuan, Wang Haiyan and Ren Xunyi, Implementation and Application of Computing Self-Similar Parameter by R/S Approach to Analyze Network Traffic, Journal of Nanjing University of Aeronautics & Astronautics, Vol.39, No.3, pp.358-362

[13] Crovella M. E. and Bestavros A., Explaining World Wide Web Traffic Self-Similarity, Technical Report TR-95-015,1995.

[14] S. Burklen, P. J. Marron, S. Fritsch and K. Rothermel, User Centric Walk: An Integrated Approach for Modeling the Browsing Behavior of Users on the Web, Annual Simulation Symposium - ANSS , 2005, pp.149-159.

[15] J. Nielsen, How Users Read on the Web, http://www.useit.com/alertbox/9710a.html, 2012.2.

[16] M. E. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic: evidence and possible causes, ACM SIGMETRICS Performance Evaluation Review, Vol.24, No.1, pp.160-169.

[17] P. Barford and M. Crovella, Generating Representative Web Workloads for Network and Server Performance Evaluation, In Measurement and Modeling of Computer Systems, 1998, pp. 151-160.

[18] J. Nielsen, How Little Do Users Read?, http://www.useit.com/alertbox/percent-text-read.html, 2012.2

[19] Xiaoliang Zhao, Daniel Massey, Mohit Lad, and Lixia Zhang. On/Off model: A new tool to understand BGP update bursts. Technical report,USC-CSD, August 2004.

[20] Dimitrios Lymberopoulos, O. Riva, K. Strauss, A. Mittal and A.Ntoulas,http://www.cs.washington.edu/homes/kstrauss/publications/asplos254-lymberopoulos.pdf, 2012.2.

[21] Chuan Xu, Mei Wang, Hong Tand., Analysis on User Click Behavior in the Mobile Internet,International Journal of Digital Content Technology and its Applications, Vol.5, No.6,pp.16-23.

 ISBN: 978-1-61208-203-5

APPENDIX

Table. Basic Statistics of Data Sets.

Date	*Number of Records*	*WAP Requests*	*WEB Requests*	*WAP Traffic Volume(GB)*	*WAP Percent*	*WEB Traffic Volume(GB)*	*WEB Percent*
2010-4-5	2275407	806752	382918	5.96	31.49%	12.98	68.51%
2010-4-6	2487533	840858	471434	6.09	30.52%	13.87	69.48%
2010-4-7	2594647	888792	560126	6.29	34.04%	12.19	65.96%
2010-4-8	2534699	846169	567362	6.37	34.14%	12.29	65.86%
2010-4-9	2578779	865051	467761	6.59	37.05%	11.19	62.95%
2010-4-10	2427638	838695	409063	6.21	31.61%	13.45	68.39%
2010-4-11	2417913	876233	408261	6.27	36.05%	11.89	63.95%
2011-4-4	17321932	5371434	11950498	37.54	23.66%	121.14	76.34%
2011-4-5	16040697	4917835	11122862	34.03	24.02%	107.67	75.98%
2011-4-6	16694404	5017418	11676986	33.33	21.76%	119.89	78.24%
2011-4-7	17838468	5435136	12403332	37.78	22.62%	129.27	77.38%
2011-4-8	17591404	5330330	12261074	36.72	22.16%	129.03	77.84%
2011-4-9	16991844	5147989	11843855	35.28	21.82%	126.45	78.18%
2011-4-10	16450519	4933624	11516895	33.59	22.13%	118.21	77.87%
The Week in 2010	17316616	5962550	3266925	43.79	33.46%	87.08	66.54%
The Week in 2011	118929268	36153766	82775502	248.3	22.57%	851.68	77.43%

 ISBN: 978-1-61208-203-5

FPGA Implementation of CRC with Error Correction

Wael M El-Medany

Computer Engineering Department,
College of Information Technology,
University Of Bahrain, 32038 Bahrain,
Email: welmedany@uob.edu.bh

***Abstract* - This paper presents a Cyclic Redundancy Check (CRC) soft core design and its hardware implementation on Field Programmable Gate Array (FPGA). The core design includes both of the Encoder and Decoder systems to be used for the serial data transmission and reception of the Wireless Transceiver System. VHDL (VHSIC Hardware Description Language) has been used for describing the hardware of the Intellectual Property (IP) core chip. The core design has been simulated using and tested using ISim (VHDL/Verilog). Spartan 3A FPGA starter kit from Xilinx has been used for downloading the design into Xilinx Spartan 3A FPGA chip.**

Keywords-FPGA; CRC Code; IP Core; VLSI.

I. INTRODUCTION

In digital communication systems, the error detection is performed by computing checksum on the message that needs to be transmitted. The computed checksum is then concatenated to the end of the message to generate the codeword or the check sequence number to be transmitted. At the receiving end, the received word is compared with the transmitted codeword. If both are equal, then the message received is treated as error free, otherwise there is an error detected in the received word.

Cyclic Redundancy Check (CRC) Code has a wide range of applications in data communications and storage devices [1-6]. Cyclic Redundancy Check (CRC) is an error-checking block code that has been used for error detection only in which the received word has to be divided by a predetermined number called the generator number. If the reminder is zero, this means that there is no error detected, for nonzero reminder, this means that there is an error detected [7-10].

Cyclic Redundancy Check (CRCs) codes are so called because the check (data verification) code is a redundancy (it adds zero information) and the algorithm is based on cyclic codes [11]. CRC has applications also in Integrated Circuits Testing Design (ICTD), and Logical Fault Detections (LFD) [12]. In [3], Albertengo et al. derived a method for determining the logic equations for any generator polynomial. Their formalization is based on z-transform. To obtain logic equations, many polynomial divisions are needed. Thus, it is not possible to write a VHDL code that generates automatically the equations for CRC.

Normally, the design of the error control decoder is more complex than the encoder. CRC when first introduced was for error detection only, it can detect single bit error; burst error with length "w", where "w" equal to the number of bits for the Frame Check Sequence (FCS) number; and odd number of errors based on the value for the generator umber [14-15]. Further research has investigated theoretically a CRC with error correction capabilities. Shukla and Bergmann [1] failed to show their hardware implementation for CRC with one bit error correction, and the simulation for correcting the bit error.

The work presented in this paper describes the VHDL implementation of a CRC Decoder that has the advantages of correcting more than one bit error. Since we are introducing the hardware implementation for error CRC with error correction, our main concern is about the design of the CRC decoder with error correcting capabilities. The error correction in CRC decoder based on the error trapping technique, which is a cyclic linear block code [16-19]. Error trapping based on, cyclic shifting the received word on the division circuit, until the error can be trapped on the parity check bits. In that case the reminder will used as the error pattern, by which we can locate and correct the detected error.

The VHDL source code has been edited and synthesized using Xilinx ISE 13.1, and then simulated and tested using ISim (VHDL/Verilog). Spartan 3A FPGA starter kit from Xilinx has been used for downloading the design into Xilinx Spartan 3A FPGA chip. The design has been tested in a hardware environment for different data inputs.

The materials in this article are organized as follows: in Section II, a brief description of the State Machine (SM) chart of CRC encoding process; the SM chart for decoding algorithm is given in Section III; the modification in the decoding algorithm for error correction will be concluded in Section IV; in Section V, the circuit design for CRC decoder will be described, as well as the top-level design the decoder; the simulation results and discussion is given in Section VI; at the end, a conclusion will be given in Section VII.

II. SM CHART FOR CRC ENCODING PROCESS

The encoder generates an n-bit check sequence number from the given input k-bit information. The encoding process starts by calculation the Frame Check Sequence (FCS), by dividing the information bits by the predefined generator

 ISBN: 978-1-61208-203-5

number. The encoder then concatenates the FSC number to the k-bit information number to get the check sequence number with n-bits length. Figure 1 shows the SM chart for the encoding algorithm. Where m is k information bits, p is the generator number, and c is the check sequence number. In Figure 1, d and r are internal signals in VHDL architecture, where d has to be divided by r in order the get the reminder, which will be used as the FCS. The division process used in the encoding algorithm is the parallel division, which will be faster than the serial one. The serial division requires a number of clock cycles equal to the number of information bits in order to calculate the FCS; however the parallel one requires only one clock cycle.

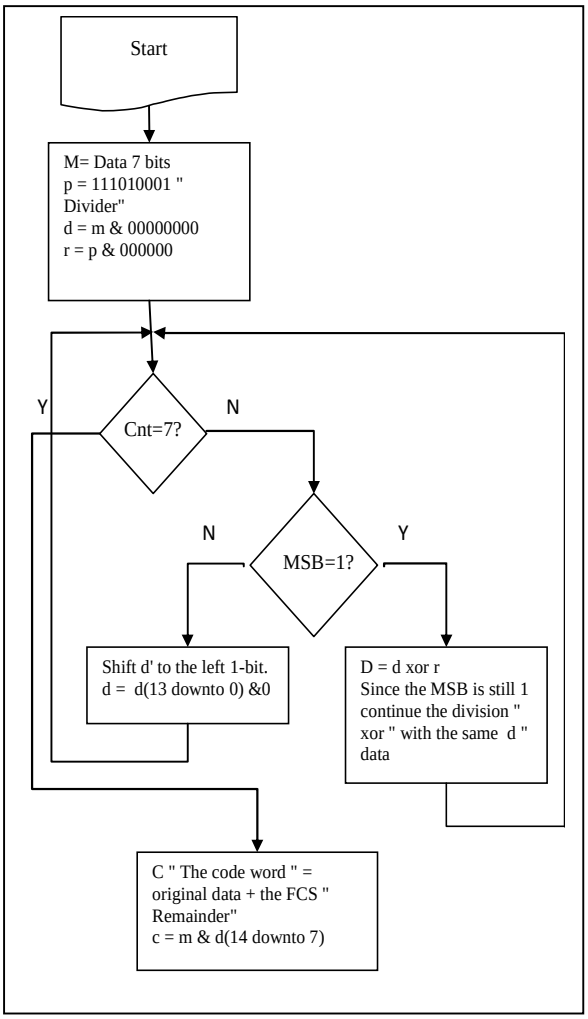


Figure 1. SM chart for CRC Encoding algorithm

III. SM CHART FOR CRC DECODING PROCESS

The CRC decoder is working similar to the encoder, where both of them based on using a division circuit. The decoding process starts by dividing the received check sequence number by the generator number. If the remainder is zero, assume there is no error detected. For nonzero remainder, it means that the received check sequence number got an error detected. The CRC decoder stop at the stage of detecting whether there is an error detected or not. But for our algorithm we are going to continue the division process until we get the error trapped, or get a decision that there is no error detected. As shown in Figure 2, the SM chart of the CRC decoding algorithm is similar to the one in Figure 1. However the decoder divide the received check sequence number 'c' by the generator number 'p', and check the reminder value whether it is zero or not, where zero reminders mean that there is no error detected, and none zero reminder means that there is an error detected.

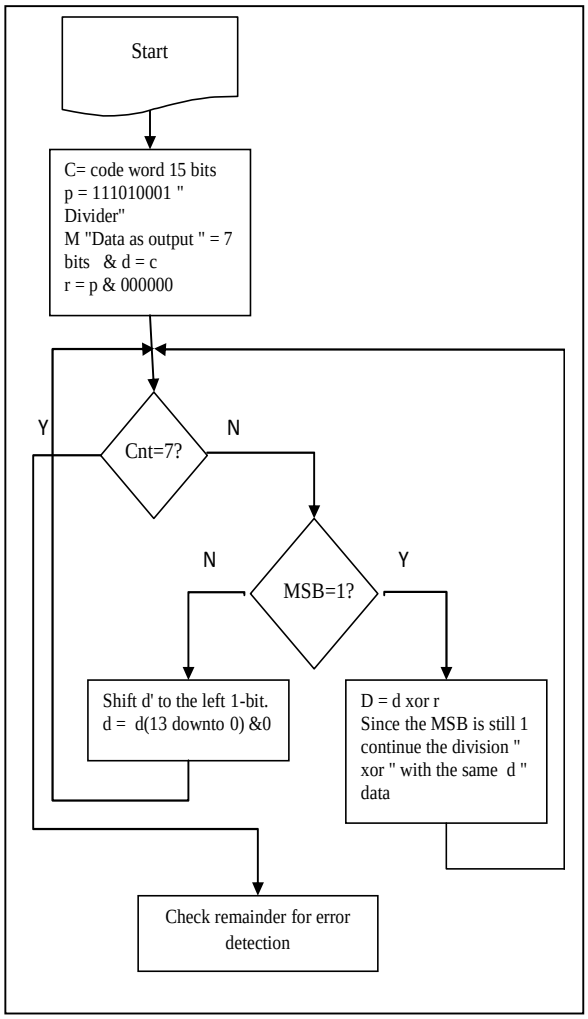


Figure 2. SM chart for CRC Decoding algorithm

IV. MODIFIED CRC DECODING ALGORITHM FOR ERROR CORRECTION

Cyclic Redundancy Check is a class of cyclic coding, which is one of the most powerful linear block codes. The modification for the CRC decoding algorithm based on the well known Error Trapping Technique (ETT), in which the

 ISBN: 978-1-61208-203-5

error has to be trapped in the parity bits. In CRC we will continue the division process until we get the error trapped in the FCS bits, or get a message that there is no error detected. This process can be done by cyclic shifting the generator number during the division process, and each time we do the division, we compare the reminder with number of errors that can be corrected using the linear block coding techniques, by calculating the Hamming distance for the code. Then use the following very famous equation:

$t = (d_{min} - 1)/2$

where t is the number of errors that can be corrected, and d_{min} is the minimum Hamming distance. The SM chart of the modified algorithm is shown in Figure 3, where the process of cyclic shifting has been added to the previous one shown in Figure 2, as well as checking the reminder after each division and compare it to the value of 't'.

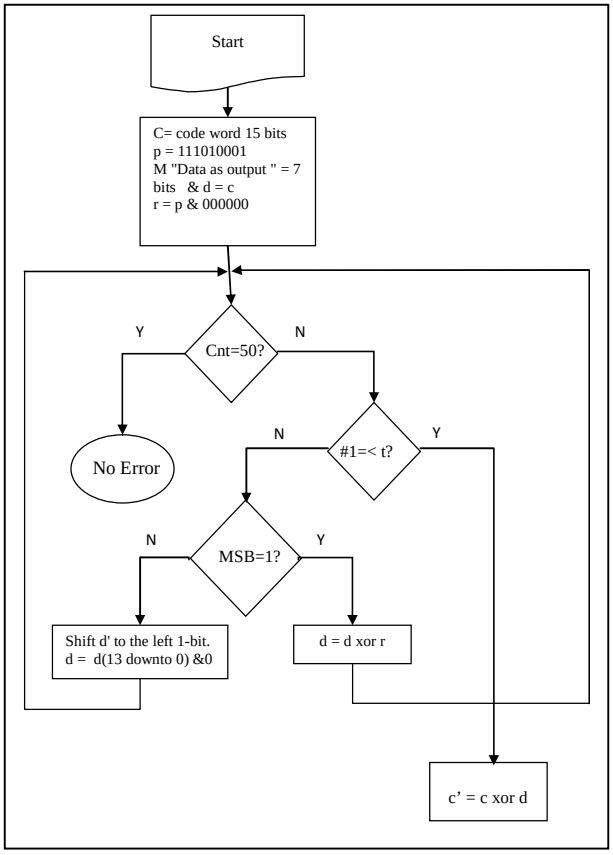


Figure 3. SM chart for a Modified CRC Decoding algorithm

If the number of non zero elements in the reminder is equal to or less than the value of 't', then the error has been trapped, and the reminder becomes the error pattern, which can be easily corrected. If this process would have been repeated a number of times, and the process has been entered in an infinite loop, then the process has to fished with the decision of "no error detected".

V. CRC Decoder Circuit Design

In this section, we are going to describe the hardware design for the CRC decoder circuit with error correction capability. The top-level design of the decoder circuit is given in Figure 4, which shows that the decoder has [Rx], in our example the received check sequence number [Rx] is 15-bit, which represent the code length, and the output of the decoder has 7-bits, which represent the information bits. In Figure 5, the second level of the top-down design is shown, the second level shows that there are three main units in the decoder circuit; the first unit is the Error Detection (ED) unit; the second unit is Locating Error (LE) unit; and the third one is the Error Correction (EC) unit. The top level of ED unit is given in Figure 6, which shows the inputs to this unit is [Rx], and the output is [sy] the syndrome that represents the remainder of the division process, and it is 8-bits, based on the values of the syndrome, whether it is zero or non zero, that will give indication whether there is an error detected or not.

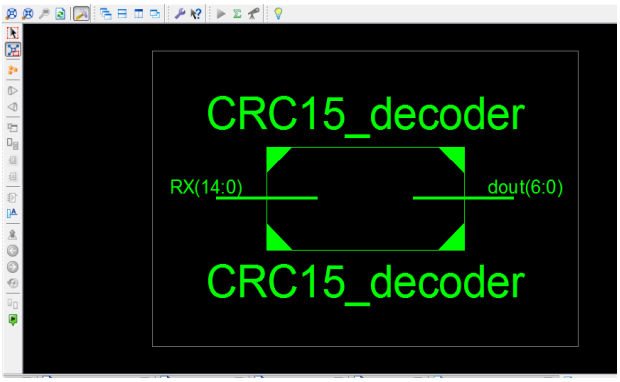


Figure 4. Top-level design of the decoder circuit

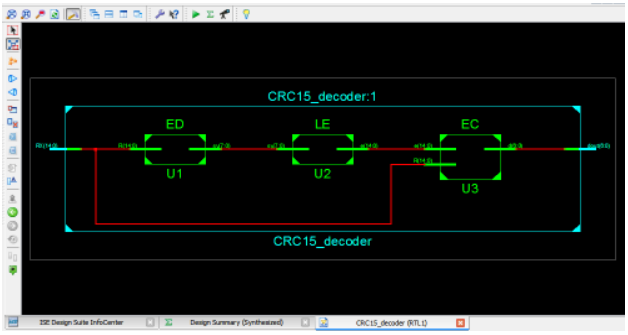


Figure 5. Second-level design of the decoder circuit

The register transfer logic for ED unit is given in Figure 7, in which there are 13 subunits, some of these subunits, we can see the gate level schematic, and some others are Xilinx FPGA building blocks. The top level of LE unit is given in Figure 8, which shows the inputs to this unit is [sy] coming from ED unit, and the output is [e] that represents the location of the bit in error. The register transfer logic for LE unit is given in Figure 9, which got only one building block

unit, called Mrom_e1, it is a ROM memory unit of the FPGA building blocks.

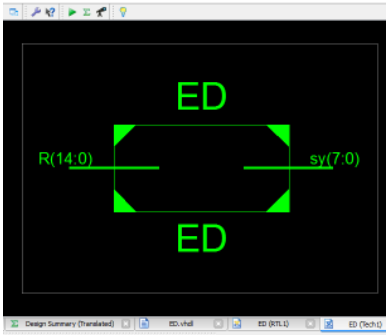


Figure 6. Top Level of the ED unit

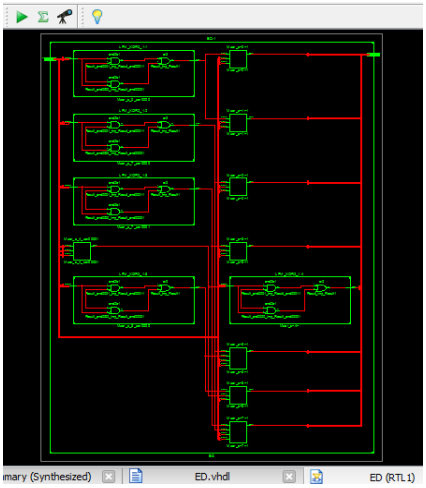


Figure 7. Register Transfer Level of the ED unit

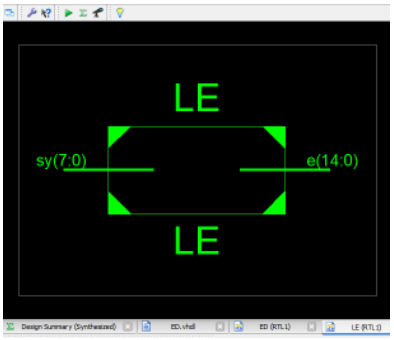


Figure 8. Top Level of the LE unit

The top level of EC unit is given in Figure 10, which shows the inputs to this unit is [R and e], and the output is [d] that represents the corrected data. The register transfer logic for EC unit is given in Figure 11, which got 7 building block units.

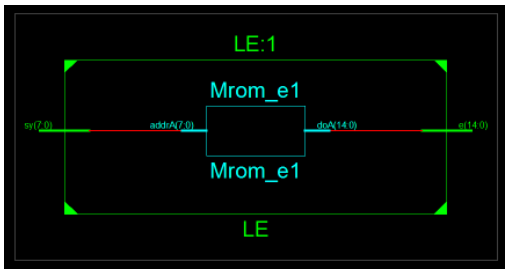


Figure 9. Register Transfer Level of LE unit

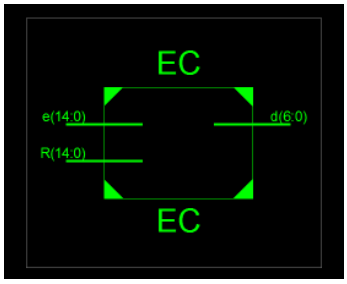


Figure 10. Top Level of the EC unit

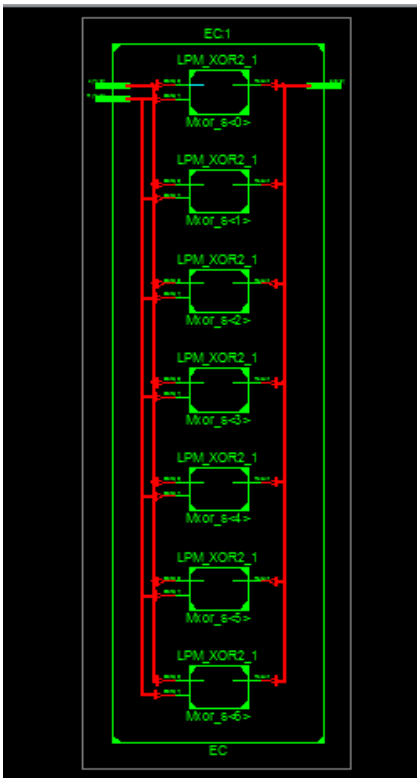


Figure 11. Register Transfer Level of the EC unit

VI. Simulation Results and Discussion

This section presents the results obtained through the simulation and design implementation summary are described. The presented CRC decoder can correct up to two errors. The gate level design of the error detection unit is given in Figure 12, and the gate level design of the error

correction unit is given in Figure 13. The test bench waveform for CRC decoder is given in Figure 14, with different inputs, and different errors. The simulation shown in Figure 14 for the modified CRC decoder and it can correct single bit error and double bit errors. For simplicity, we are given errors two codewords all one's codeword and all zero's codeword, "11111 11111 11111" and "00000 00000 00000". The inverted bits represent the introduced error, which is also represented by the signal [s2]. The signal [dout] is 7-bit corrected information, the simulation results for the presented CRC decoder proved the correctness of the decoder circuit either for error detection or error correction.

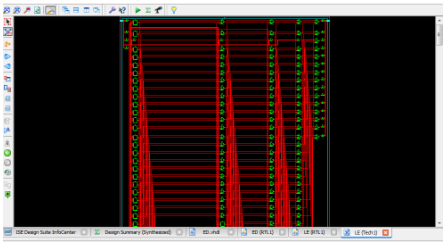


Figure 12. Gate Level Design of the ED unit

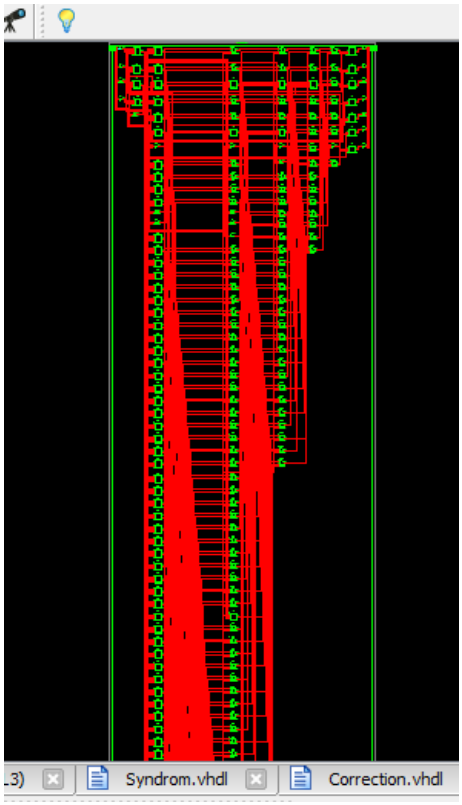


Figure 13. Gate Level Design of the EC unit

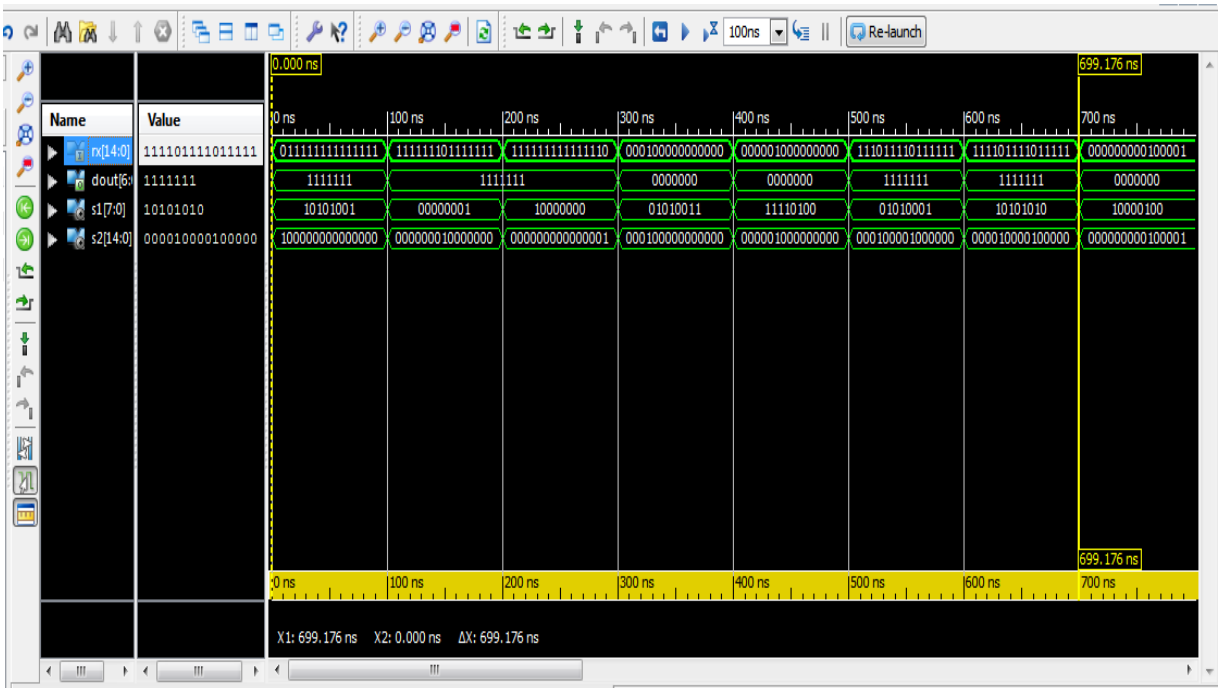


Figure 14. Test bench for CRC Decoder

VII. CONCLUSIONS

An FPGA design of a CRC decoder with error correction capabilities have been simulated and implemented. The system has been designed using VHDL, and implemented on hardware using Xilinx Spartan 3AN FPGA Starter kit. The CRC decoder for both error detection and error correction have been tested for different data inputs either for simulation purposes or in the hardware environment using the available FPGA kit. The VHDL source code has been edited and synthesized using Xilinx ISE 13.1, and then simulated and tested using ISim (VHDL/Verilog).

REFERENCES

[1] Shukla S. and Bergmann N. W., "Single bit error correction implementation in CRC-16 on FPGA", 1n: I EEE International Conference on Field-Programmable Technology. Brisbane, Australia, 2004. .

[2] W.W.Peterson and D.T.Brown, "Cyclic Codes for Error Detection", *Proc. IRE*, Jan. 1961.

[3] G. Albertengo and R. Sisto, "Parallel CRC Generation", *IEEE Micro*, Oct. 1990.

[4] R. Lee, "Cyclic Codes Redundancy," *Digital Design*, July 1977.

[5] A. Perez, "Byte-wise CRC Calculations", *IEEE Micro*, June 1983.

[6] A. K. Pandeya and T. J. Cassa, "Parallel CRC Lets Many Lines Use One Circuit", *Computer Design*, Sept. 1975.

[7] A. S. Tanenbaum, "*Computer Networks"*, Prentice Hall, 1981.

[8] T. V. Ramabadran and S. S. Gaitonde, "A tutorial on CRC computations", *IEEE Micro*, Aug. 1988.

[9] W. W. Peterson and D. T. Brown, "Cyclic Codes for Error Detection", *Proc. IRE*, Jan. 1961.

[10] M. Sprachmann, "Automatic generation of parallel CRC circuits", IEEE Des. Test Comput., vol. 18, no. 3, pp. 108–114, May/Jun. 2001.

[11] N.R.Sexana and E.J.McCluskey, "Analysis of Checksums, Extended Precision Checksums and Cyclic Redundancy Checks", *IEEE Transactions on Computers*, July 1990.

[12] J.McCluskey, "High Speed Calculation of Cyclic Redundancy Codes," in *Proc. of the 1999 ACM/SIGDA seventh Int. Symp. on Field, 1999.*

[13] K. V. GANESH, D. SRI HARI, and M. HEMA, "Design and Synthesis of a Field Programmable CRC Circuit Architecture", International Journal of Engineering Research and Applications, ISSN 2248-9622**,** Volume 1, Issue 4, Nov-Dec 2011.

[14] William Stallins, "Data and Computer Communications", Eight edition, Prentice Hall, 2007.

[15] Behrouz A. Forouzan, "Data Communications and Networking", third edition, McGraw Hill, 2003.

[16] S. Lin and D. J. Cosetllo, "Error Control Coding: Fundamentals and Applications", Prentice Hall, NJ, 1983.

[17] R.E. Blahut, "Theory and Practice of Error Control Codes", Addison-Wesley, Menlo Park, California, 1983.

[18] G. Campobello, M. Russo, and G. Patanè, "Parallel CRC realization", IEEE Trans. Comput., vol. 52, no. 10, pp. 1312–1319, Oct. 2003.

[19] *Programmable Gate Arrays*, p. 250, ACM Press New York, NY, USA, 1999.

[20] G. Sharma‡, A. Dholakia□, and A. Hassan, "Simulation of Error Trapping Decoders on a Fading Channel", Proc. IEEE Vehicular Technology Conference, Atlanta, GA, 28 Apr.-1 May 1996, vol. 2, pp. 1361-1365

[21] M. J. S. Smith, "*Application-Specific Integrated Circuits"*, Addison-Wesley Longman, Jan. 1998.

[22] P.C. Hershey and C. B. Silio, "Finite State Machines for Information Collection and Assessment on High Speed Data Networks", Wireless and Optical Communications Proceeding, 2002.

[23] C. Borrelli, "IEEE 802.3 Cyclic Redundancy Check", application note: Virtex Series and Virtex-II Family, XAPP209 (v1.0), Xilinx, Inc, March 23, 2001.

[24] Efficient LDPC Decoder Implementation for DVB-S2 System, Apr 2010.

[25] D. Giot, P. Roche, G. Gasiot, and R. Harboe-Sorensen, "Multiple-Bit Upset Analysis in 90 nm SRAMs: Heavy Ions Testing and 3D Simulations", IEEE Trans. Nucl. Sci., Vol. 54, pp.904 – 911, Aug. 2007.

[26] Xilinx, "Spartan-3E Starter Kit Board User Guide", Xilinx, Tech. Rep. UG230, Mar 2006.

[27] Xilinx, "Virtex 5 Family Overview", Xilinx, Tech. Rep. DS100, Jun 2008.

 ISBN: 978-1-61208-203-5

On Throughput Characteristics of Type II Hybrid-ARQ with Decode and Forward Relay using Non-Binary Rate-Compatible Punctured LDPC Codes

Hironori Tanaka
Dept. of Computer Science and Engineering
Nagoya Institute of Technology
Nagoya, Japan
E-mail: 22417569@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, an NB RCP LDPC (Non-Binary Rate-Compatible-Punctured Low Density Parity Check) code is designed over the extended Galois Field. The designed NB RCP LDPC code is applied to the type II HARQ (Hybrid Automatic Repeat reQuest) with Decode and Forward (DF) relay using MIMO-OFDM modulation. The designed code enables us to decrease the coding rate with incremental redundancy for each retransmission in HARQ. The retransmission is made by the DF relay after its successful decoding. We have verified through computer simulations that the proposed type II HARQ scheme with DF relay greatly improves the throughput and average retransmission characteristics compared with the scheme without DF relay.**

Keywords-NB RCP LDPC code; Hybrid-ARQ; Decode and Forward Relay; MIMO-OFDM; Symbol-LLR.

I. INTRODUCTION

An LDPC code which suits the flexible coding rate design and has the high error correcting capability through iterative decoding can be constructed on arbitrary extended Galois field. The Non-Binary (NB) LDPC code constructed on extended Galois field generally exhibits the better BER performance than the binary LDPC codes [1],[2]. There exist also Rate-Compatible-Punctured (RCP) LDPC codes with variable coding rate obtained by properly puncturing the mother LDPC code. The RCP LDPC codes enable us to use the same decoder as the mother code [3] and suit the ARQ (Automatic Repeat reQuest) error correcting schemes [4],[5] with the incremental redundancy. When comparing the HARQ using NB RCP LDPC codes with the existing RCPT (Rate Compatible Punctured Turbo) HARQ using binary Turbo codes [6], the HARQ with NB LDPC codes can cope with flexible coding rates, code word lengths and NB symbol LLR additions without using inter-leavers for burst errors on the channel. By combining the NB LDPC codes with the RCP codes, the NB RCP LDPC codes were designed and the designed NB RCP LDPC codes were applied to the type II HARQ [7],[8]. In this paper, the NB RCP LDPC coded type II HARQ with the MIMO-OFDM modulation is used for the Decode and Forward (DF) relay scheme [9],[10]. By using the DF relay, the source node can be replaced by the relay, once the relay correctly decodes the LDPC encoded packet from the source. This replacement from the source to the relay effectively reduces the number of retransmissions and improves the throughput performance very much. We have verified through computer simulations that the proposed DF relaying scheme with type II HARQ and RCP LDPC code greatly improves the throughput and average retransmission characteristics compared with the case without DF relay.

The paper is organized as follows. In Section II, the RCP LDPC code is introduced. In Section III, NB LDPC coded Type II HARQ scheme is described. In Section IV, we propose decode and forward relaying scheme. In Section V, we present the symbol LLR generation in OFDM demodulation. In Section VI, the computer simulation results are shown. The paper concludes with Section VII.

II. RCP LDPC CODE

The encoding and decoding procedure of RCP LDPC code is as follows. We call the code before puncture and the code after puncture as the mother code and the efficient code, respectively. In RCP LDPC code, the encoder and decoder of mother code can also be applied to the efficient code. When the parity check matrix of mother code is given by $\boldsymbol{H}_M\,(M \times N)$ and the generator matrix by $\boldsymbol{G}_M\,(N \times K)$ with $K = (N - M)$, the coding rate of mother code becomes $R_M = (1 - M / N) = K / N$. The coding rate after the puncture of p symbols from the mother code is given by $R_E = K/(N - P)$. We denote the message vector as $\boldsymbol{m} = (m_1, m_2, \ldots, m_K)$, the code word of mother code as $\boldsymbol{C_M} = (C_{M1}, C_{M2}, \ldots, C_{MN})$, the index of position to be punctured as $\boldsymbol{P} = (p_1, p_2, \ldots, p_p)$ and the code word of efficient code as $\boldsymbol{C}_E = (C_{E1}, C_{E2}, \cdots, C_{EN})$. The encoding procedure is first to generate the mother code by $\boldsymbol{C_M} = \boldsymbol{m}\boldsymbol{G_M}$ and next to puncture the position using $\boldsymbol{P}$ to obtain $\boldsymbol{C}_E$. The decoding procedure is to produce the symbol LLR from the receive signal and it is fed to the mother code decoder as the initial value for the sum-product algorithm. The symbol LLR for the position $\boldsymbol{P}$ is initially set to 0, because there is no available symbol LLR corresponding to the position $\boldsymbol{P}$.

III. NB RCP LDPC CODED TYPE II HARQ SCHEME

In Fig. 1, we show the block diagram of NB RCP LDPC coded Type II HARQ scheme using MIMO-OFDM modulation. At the transmitter, the data bits are firstly encoded by the CRC-16 error detecting code and secondly encoded by the NB LDPC code on GF(4) or GF(16). The encoded LDPC code word is divided into the OFDM frames for making a packet for each transmission using the predetermined puncture table introduced in [3]. In Fig. 2, we show how to divide the coded symbols of an LDPC

 ISBN: 978-1-61208-203-5

code to the OFDM frames. The encoded NB alphabets are mapped to QPSK signal points for GF(4) or 16QAM for GF(16). These signal points are then modulated by OFDM with guard interval insertion. The OFDM signal is then transmitted to the quasi-static frequency selective channel from each transmit antenna. At the receiver, for each antenna, the guard interval is first removed and then OFDM demodulation is made using FFT. By demodulating each subcarrier QPSK modulated or 16QAM modulated, the symbol LLR (Log Likelihood Ratio) is calculated. The symbol LLR will be defined in Section V. The symbol LLR values are then fed to the LDPC decoder and the iterative decoding using sum-product algorithm is made. The decoded information bits are error-detected by the CRC-16 code. If error is not detected, the data bits are fed to the data sink and the ACK is returned to the source node (transmitter) to finish the transmission. But if errors are detected, the NACK is returned and the retransmission is requested. As the type II Hybrid ARQ (HARQ) scheme is employed, at the first transmission, only the data symbols without encoding are sent to the receiver. After the 2nd transmission, as shown in Fig. 2, the parity symbols are sent several times with the incremental redundancy depending on the error detection status at the receiver. When the channel quality is good, the uncoded data packet for the first transmission succeeds with high probability leading to the high throughput performance. On the other hand, when the channel quality is bad, the parity packets are retransmitted several times till the LDPC code rate reaches the lowest one half resulting in enough error correction capability.

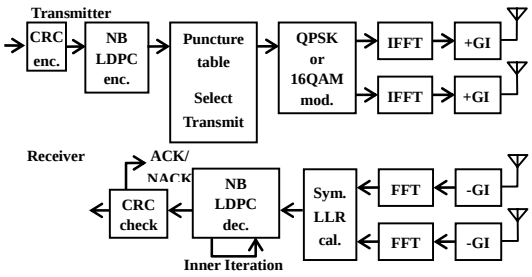


Fig. 1 Transmitter and receiver configuration of MIMO NB RCP LDPC coded type II HARQ scheme

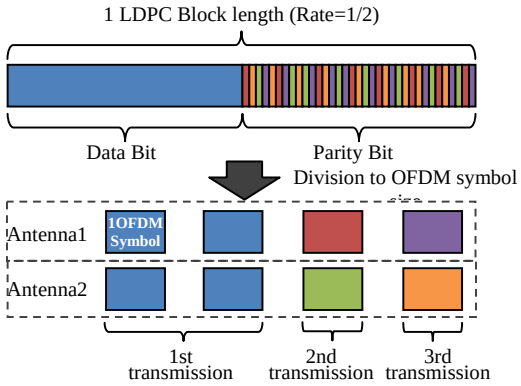


Fig. 2 OFDM frames obtained from the division of an LDPC code word

At the transmitter side, only one time of encoding process of RCP LDPC code is enough and there is no need of re-encoding process for decreasing the coding rate thereafter. Accordingly the complexity of encoding process of RCP LDPC code does not increase compared with the fixed code rate LDPC code. At the receiver side, the same decoder can be used for each coding rate, so there is no increase of complexity when compared with the fixed coding rate.

IV. DECODE AND FORWARD RELAYING SCHEME

The Decode and Forward relay model is shown in Fig. 3 and is composed of source, relay and destination. In this model, we assume the relay is located at the middle point on the straight line between the source and destination. When the receive power at the receiver attenuates in proportion to $1/d^{\alpha}$ where d is the distance between transmitter and receiver, the receive power from source to relay and the one from relay to destination become 2^{α} times larger than the one from source to destination. Next, we will illustrate the operation at each node in Fig. 3. At the source node, the encoding and modulation process using the NB RCP LDPC coded type II HARQ with MIMO OFDM is made, and uncoded and parity check packets are generated using the division of an LDPC code word as shown in Fig. 2. At the 1st transmission, the source broadcasts the uncoded data packet to the destination and the relay simultaneously. The relay and the destination receive the packet and they make the error detection independently using CRC-16 code. Then the destination and relay send the ACK or NACK back to the source, and the error detection results are shared among the source, relay and destination. When the destination returns ACK, the data are received successfully at the destination through only one transmission and this situation is equivalent to the case without relay. On the other hand, when the destination returns NACK, retransmission must be made. Moreover, when the relay also returns NACK, the parity check packet with incremental redundancy is retransmitted from the source. At the destination and the relay, the received parity check packet is combined with the data packet for the 1st transmission, and the combined LDPC code word is decoded. Then the decoded information bits are CRC checked and ACK or NACK is returned to the source. If the destination returns NACK but the relay dose ACK, then the relay receives the correct information bits. Accordingly, the relay can encode the received information bits to the same RCP LDPC code word with the source, i.e., we can replace the source by the relay for the subsequent ARQ retransmission. The parity check packet with incremental redundancy is generated at the relay and is sent from the relay afterwards. Once the relay successfully receives the packet from the source, then the relay performs the function of source and the source stops any further retransmission. This strategy is quite useful because the relay is closer to the destination than the source, thus the transmission error does not occur so frequently compared with the source.

 ISBN: 978-1-61208-203-5

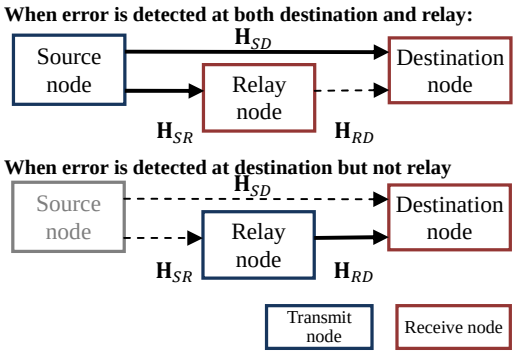


Fig. 3 DF (Decode and Forward) relaying model

Another point to be noticed is that the consumption of total transmit power remains the same as the one without the relay, because the source is replaced by the relay and the source does not consume any transmission power after the replacement.

V. SYMBOL LLR GENERATION IN OFDM DEMODULATION

As an example of symbol LLR calculation in OFDM demodulation, we show the case where the NB LDPC code on GF(4) is used and the QPSK modulation is employed for each subcarrier of OFDM. When the transmit signal, receive signal, signal points of QPSK and the subcarrier channel fading value are denoted as x, r, s_0, s_1, s_2, s_3 and h respectively, the symbol LLR for the alphabets $a = 0,1,2,3$ on GF(4) is defined as

$$LLR_a = \log_e\left\{\frac{P(x=a|r\Delta r)}{P(x=0|r\Delta r)}\right\} = \log_e\left\{\frac{P(s_a, r\Delta r)/P(r\Delta r)}{P(s_0, r\Delta r)/P(r\Delta r)}\right\}$$
$$= \log_e\left\{\frac{P(s_a, r\Delta r)}{P(s_0, r\Delta r)}\right\} = \log_e\left\{\frac{P(s_a)\,p(r\Delta r|s_a)}{P(s_0)P(r\Delta r|s_0)}\right\} \quad (1)$$
$$= \log_e\left\{\frac{p(r|s_a)\Delta r}{P(r|s_0)\Delta r}\right\} = \log_e\frac{p(r|s_a)}{P(r|s_0)}$$

where the priori probabilities are set to $P(s_0) = P(s_1) = P(s_2) = P(s_3) = 1/4$, i.e., equal probabilities. In (1), $P(x=a|r\Delta r)$ denotes the probability that the transmit symbol x equals a when the receive signal point r falls in the small area $r\Delta r$ centered at r. $p(r|s_a)$ is the transition probability density function from $s_a \to r$ and is expressed as

$$p(r|s_a) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{|r-hs_a|^2}{2\sigma^2}\right) \quad (2)$$

Accordingly, the symbol LLR is calculated as

$$LLR_a = \log_e\left\{\frac{p(r|s_a)}{p(r|s_0)}\right\} = \log_e\left\{\frac{\frac{1}{2\pi\sigma^2}\exp\left(-\frac{|r-hs_a|^2}{2\sigma^2}\right)}{\frac{1}{2\pi\sigma^2}\exp\left(-\frac{|r-hs_0|^2}{2\sigma^2}\right)}\right\}$$
$$= \left(-\frac{|r-hs_a|^2}{2\sigma^2}\right) - \left(-\frac{|r-hs_0|^2}{2\sigma^2}\right) = \frac{|r-hs_0|^2 - |r-hs_a|^2}{2\sigma^2} \quad (3)$$

VI. COMPUTER SIMULATION RESULTS

The BER characteristics of RCP LDPC code on AWGN channel are examined when the rate 1/2 mother code on GF(4) or GF(16) is punctured to change the coding rate. The simulation condition is listed in Table I and the simulation results are shown in Figs. 4 and 5. From the simulation results, we know that the efficient codes on GF(4) or GF(16) with different coding rates are obtained from a mother code and the error correction capability corresponding to each coding rate is achieved. Next, the throughput performance and the average number of retransmission characteristic for 4×4 MIMO-OFDM NB RCP LDPC coded Type II Hybrid-ARQ with GF(4) and QPSK modulation are investigated. We compared the proposed relay model with the one without relay. The simulation condition is listed in Table II. The simulation results for throughput characteristic are shown in Fig. 6 and Fig. 7. The simulation results for average number of retransmission are shown Fig. 10 and Fig. 11.

TABLE I Simulation condition of RCP LDPC code

Channel		AWGN	
Modulation		QPSK	16QAM
Size of Galois field		GF(4)	GF(16)
Mother code	Size of parity check matrix	(256,512)	(128,256)
	Average weight	(2.66,5.32)	(2.41,4.82)
	Coding rate	4/8	2/4
Efficient code	Information bit length	512	
	Coding rate	4/8,4/7,4/6,4/5, 4/4	2/4,2/3,2/2
Max SPA iteration		20	

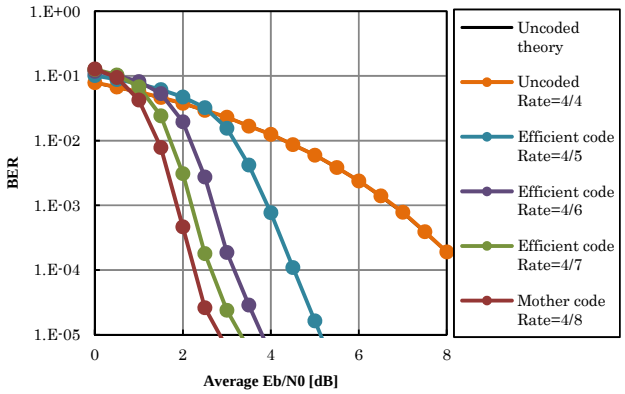


Fig. 4 BER characteristics of RCP-LDPC code on AWGN channel (QPSK)

 ISBN: 978-1-61208-203-5

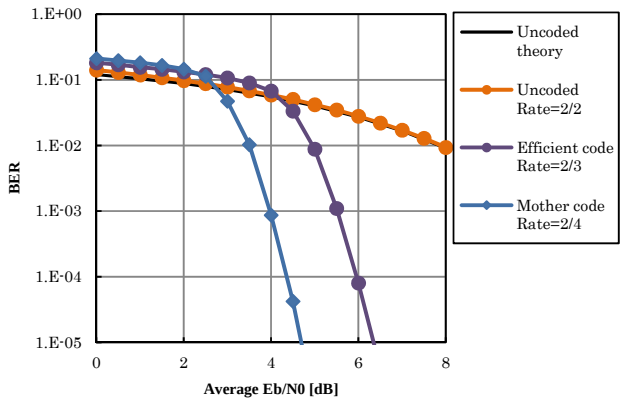


Fig. 5 BER characteristics of RCP-LDPC code on AWGN channel (16QAM)

In the simulation, an LDPC code word is divided into 32 OFDM symbols. As the coding rate of mother LDPC code is 1/2, the former 16 OFDM symbols are the information data symbols and the latter 16 OFDM symbols are the parity check symbols. For the 1st transmission, 16 OFDM symbols made from information data are transmitted from 4 transmit antennas simultaneously using 4 OFDM symbol duration. For the 2nd transmission and thereafter, i.e., retransmission, 4 OFDM symbols made from the parity check symbols are transmitted from 4 transmit antennas simultaneously. In each retransmission, 1 parity check OFDM symbol is transmitted from each antenna using 1 OFDM duration. The coding rate is decreased gradually from 4/5, 4/6, 4/7 to 4/8 for each retransmission. After all the parity check OFDM symbols are transmitted and the coding rate reaches 4/8=1/2, if the errors are still detected at the destination, the whole transmission of the same RCP LDPC code word in the same manner is repeated up to 3 times. The symbol LLR combining is used at the destination for the repeated reception of the same RCP LDPC code word. For the comparative scheme, we considered the type I HARQ with the fixed coding rate

TABLE II Simulation condition of NB GF(4) RCP LDPC coded type II Hybrid ARQ scheme with 4 × 4 MIMO-OFDM

Channel		Quasi-static equal power Rayleigh fading channel with 16 delay paths
Number of transmit and receive antennas		4×4
Power attenuation constant of channel		α = 3
Modulation		QPSK-OFDM
Size of Galois field		GF(4)
Mother code	Size of parity check matrix	(1024,2048)
	Average weights	(2.66,5.32)
	Coding rate	4/8
Efficient code	Length of information bits	2048
	Coding rates	4/4,4/5,4/6,4/7,4/8
Max SPA iteration		20
Number of OFDM subcarriers		64
GI length		16 (=*T*/4)
Delay interval		1 (=*T*/64)
Channel State Information (CSI)		Perfect at receiver
Error detection code		CRC-16 code

LDPC code and set the maximum number of repetition also to be 3. The packet combining at the destination is also used through the symbol LLR addition. Next, using the NB RCP LDPC code on GF(16) and with 16QAM modulation, we made the similar simulation to the above GF(4) with QPSK modulation. The simulation condition is given in Table III. The simulation results for throughputs are shown in Fig. 8 and Fig. 9. The simulation results for average number of retransmission are shown in Fig. 12 and Fig. 13. In these simulations, an LDPC code word is divided into 4 information data packets and 4 parity check packets. For each retransmission, the coding rate is decreased from 2/2, 2/3 to 2/4. When the all OFDM symbols are transmitted and the coding rate reaches 2/4=1/2, and if the error is still detected, the same RCP LDPC code word is repeatedly transmitted up to 3 times.

TABLE III Simulation condition of NB GF(16) RCP LDPC coded type II Hybrid ARQ scheme with 2 × 2 MIMO-OFDM

Channel		Quasi-static equal power Rayleigh fading channel with 16 delay paths
Number of transmit and receive antennas		2×2
Power attenuation constant of channel		α = 3
Modulation		16QAM-OFDM
Size of Galois field		GF(16)
Mother code	Size of parity check matrix	(128,256)
	Average weights	(2.41,4.82)
	Coding rate	2/4
Efficient code	Length of information bits	512
	Coding rates	2/2,2/3,2/4
Max SPA iteration		20
Number of OFDM subcarriers		64
GI length		16 (=*T*/4)
Delay interval		1 (=*T*/64)
Channel State Information (CSI)		Perfect at receiver
Error detection code		CRC-16 code

We compare the type II HARQ with type I HARQ in Fig. 6, 7, 8, and 9. As the type I HARQ scheme has the fixed coding rate, the throughput for each coding rate saturates to the certain value less than the maximum in high average receive E_b / N_0 region, while the throughput of type II HARQ approaches almost the maximum value of 8 (bits/sec/Hz) by adaptively changing the coding rate. The reason why the final throughput for type II HAQR is slightly less than 8 (bits/sec/Hz) is due to the use of CRC-16 code to detect the errors in information data. In type II HARQ, however, the parity check packet is sequentially retransmitted in responding to the NACK, so the number of retransmission becomes large compared with the type I HARQ. Also in type II HARQ, the iterative decoding of LDPC code is done for each retransmission of parity check packet, thus the decoding time tends to increase.

Next, we compare the cases with and without relay. When the average receive E_b / N_0 is high, the throughputs with and without relay are almost equal, but when the average receive E_b / N_0 is low, the throughput with relay is higher than without relay.

 ISBN: 978-1-61208-203-5

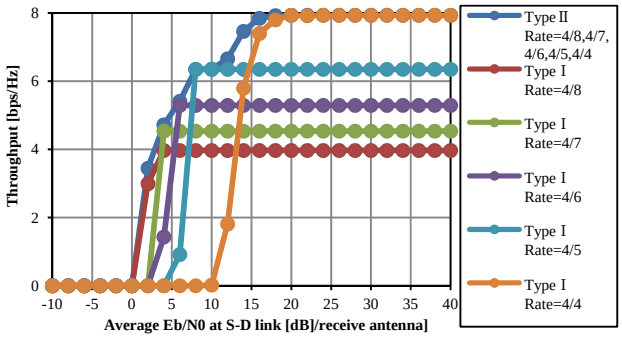


Fig. 6 Throughput characteristics of NB GF(4) LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (without relay, 4×4 , QPSK)

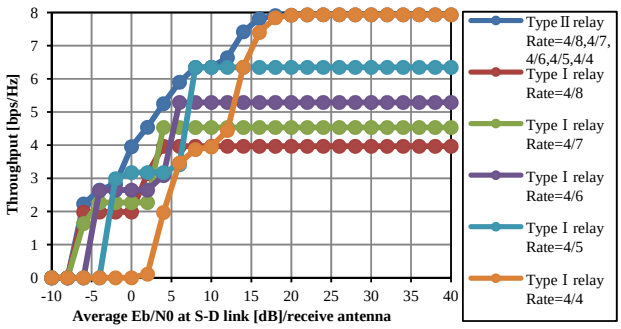


Fig. 7 Throughput characteristics of NB GF(4) LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (with relay, 4×4 , QPSK)

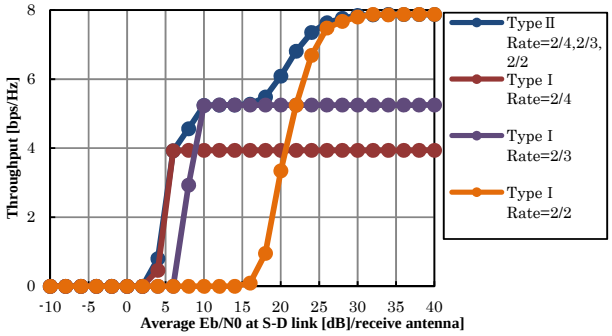


Fig. 8 Throughput characteristics of NB GF(16) LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (without relay, 2×2 , 16QAM)

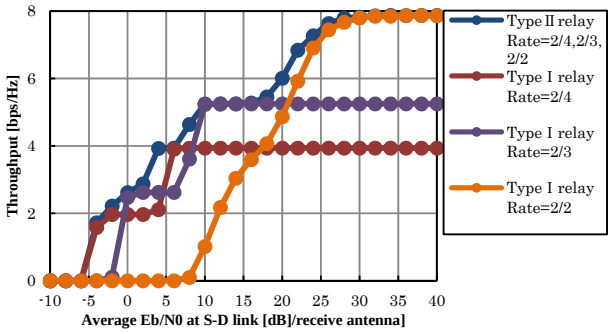


Fig. 9 Throughput characteristics of NB GF(16) LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (with relay, 2×2 , 16QAM)

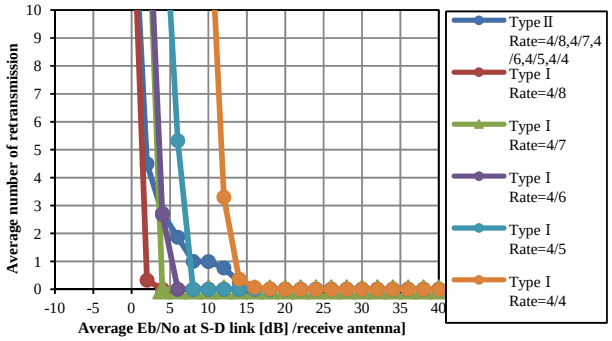


Fig. 10 Average number of retransmission of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (without relay, 4×4 , QPSK)

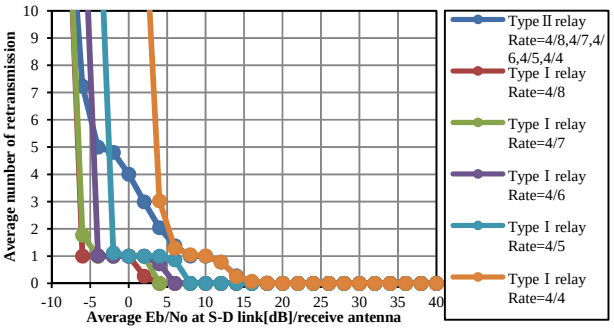


Fig. 11 Average number of retransmission of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (with relay, 4×4 , QPSK)

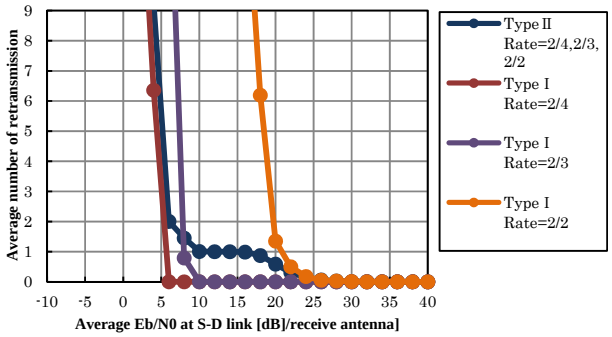


Fig. 12 Average number of retransmission of NB GF(16) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (without relay, 2×2 , 16QAM)

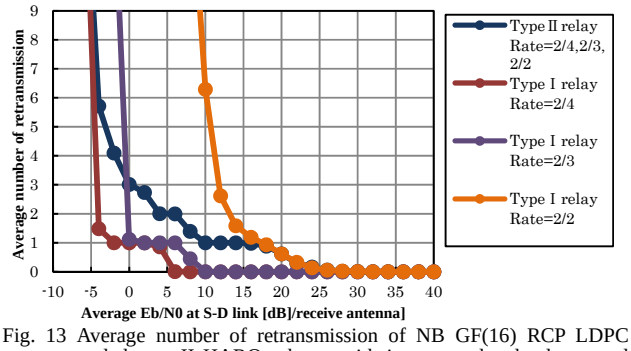


Fig. 13 Average number of retransmission of NB GF(16) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (with relay, 2×2 , 16QAM)

 ISBN: 978-1-61208-203-5

This is because for the high average receive E_b / N_0 region, the destination can receive the packet correctly without retransmission. Accordingly, the relay is not used for this high E_b / N_0 region, so there is no difference between with and without relay. On the other hand, for the region where the average receive E_b / N_0 is low, the transmission from source to destination often fails, but the transmission from relay to destination succeeds with high probability, thus the retransmission is switched from the source to the relay for this low E_b / N_0 region. For the type I HARQ schemes in Fig. 7, Fig. 11, Fig. 9 and Fig. 13, we know that the throughput with relay is largely improved compared with the one without relay for the region where the average number of retransmission is 1. For this region the throughput of type I HARQ is almost one half of the throughput for high E_b / N_0 region. This means that for this region the transmission is switched from the source to relay and the retransmission from the relay to destination is almost successful. This observation proves that the use of relay is quite effective in HARQ.

As for the proposed type II HARQ, the throughput is larger than all the type I HARQ schemes and is optimum for all average receive E_b / N_0 values. However, the average number of retransmission becomes larger than the type I ARQ schemes because of the incremental redundancy retransmissions.

VII. CONCLUSIONS AND FUTURE WORKS

In this paper, we applied the NB RCP LDPC code to the type II HARQ scheme with Decode and Forward relay. We simulated the throughput and average retransmission characteristics and showed the effectiveness of NB RCP LDPC coded type II HARQ with relay. As for the modulation scheme, we considered MIMO-OFDM with QPSK and 16QAM. Quasi-static frequency selective Rayleigh fading channel is considered between each transmit and receive antenna. The relay is located in the middle between source and destination. In the proposed type II HARQ scheme, the first transmission is made without error correcting code, i.e., the information packet with CRC check is broadcasted to the relay and the destination. If the error is detected at destination, the parity check packet is retransmitted from the source. At the relay and the destination, the information packet and the parity check packet are combined and the LDPC decoding is done. If the error is not detected at the relay, but is detected at the destination, then the relay retransmits the parity check packet in place of the destination. This means that the source is replaced by the relay. If the error is still detected at the destination, the parity check packets are retransmitted from the relay several times with the incremental redundancy till the coding rate reaches 1/2. We clarified that by using the proposed HARQ relaying scheme, the higher throughput and the fewer average number of retransmissions are achieved for the low average receive S/N region comparing to without relay.

In the future study, we will investigate the scheme in which the transmitter knows the CSI, thus the redundancy of the parity check packet can be controlled by the transmitter, leading to fewer retransmissions.

ACKNOWLEDGEMENT

This study is partially supported by the A-STEP by Japan Science and Technology Agency, and the Sharp Corporation.

REFERENCES

[1] D. Declercq and M. Fossorier, “Decoding algorithm for nonbinary LDPC codes over GF(q), ” *IEEE transactions on Communications*, Vol.55, pp. 633-643, April 2007.

[2] D. Kimura, F. Guilloud, and R. Pyndiah, “Application of non-binary LDPC codes for small packet transmission in vehicle communications,” *The 5th International Conference on ITS Telecommunications*, pp.109-112, Brest France, June 2005.

[3] J. Ha, J. Kim, D. Klinc, and S. W. McLaughlin, “Rate-compatible punctured low-density parity-check codes with short block lengths, ” *IEEE transactions on Information Theory*, vol.52, No.2, pp.728-738, Feb. 2006.

[4] Y. Tsuruta, Y. Iwanami, and E. Okamoto, “An evaluation of throughput performance of MIMO-OFDM-MLD Hybrid ARQ with bit LLR combining,” *Proc. IEICE Gen. Conf*., B-5-19, p.452, March 2009.

[5] M. Shimotsu, Y. Iwanami, and E. Okamoto, “An LDPC coded adaptive hybrid ARQ scheme with packet combining on MIMO eigen-mode channels,” *IEICE Technical Report*, RCS2005-37, pp.59-64, June 2005.

[6] D. Gang, R. Kimura, and F. Adachi, “Performance evaluation of RCPT Hybrid ARQ schemes for DS-CDMA mobile radio over frequency selective Rayleigh fading channel,” *IEICE Technical Report*, RCS2001-280, pp.241-248, March 2002.

[7] T. Kozawa, Y. Iwanami, E. Okamoto, R. Yamada, and N. Okamoto, “An evaluation on throughputs for Hybrid-ARQ using Non-Binary Rate-Compatible LDPC codes,” *The 32nd Symposium on Information Theory and its Applications (SITA2009)*, F21-3, pp.771-775, Dec. 2009.

[8] T. Kozawa, Y. Iwanami, E. Okamoto, R. Yamada, and N. Okamoto, “An evaluation on throughput performance for Type II Hybrid-ARQ using non-binary Rate-Compatible-Punctured LDPC codes,” *IEICE transactions on fundamentals*, Vol.E93, No.11, pp.2089-2091, November 2010.

[9] J. Laneman, D. Tse, and G. Wornell, “Cooperative diversity in wireless networks: efficient protocols and outage behaviour,” *IEEE transactions on Information Theory*, Vol. 50, No.12, pp.3062-3080, Dec. 2004.

[10] A. B. A. Aziz and Y. Iwanami, “A simple symbol estimation for soft information relaying in cooperative relay channels,”*Int. Journal of Commun. Networks and Systems* (IJ*CNS*), Scientific Research Publishing, Vol. 4, No. 9, pp.568-577, Sept. 2011.

 ISBN: 978-1-61208-203-5

Using Meta-Heuristic Algorithms for Minimizing the Costs of Access-Point Location

Pawel Aksiutin, Dymitr Paremski, Iwona Pozniak-Koszalka, Leszek Koszalka, Andrzej Kasprzak
Department of Systems and Computer Networks, Wroclaw University of Technology,
Wroclaw, Poland
e-mail: iwona.pozniak-koszalka@ pwr.wroc.pl

***Abstract* — Cost and quality are important issues in the network design. This paper presents an approach to wireless network cost optimization. The suggested model aims at determining optimal locations of access points which could provide service for receivers. The target service area would be determined by distribution of receivers and the access points would provide service coverage to serve prospective user-traffic demand in the selected area. Two created algorithms for location of access points are described and investigated with the designed experimentation system. It may be observed that using these algorithms in designing wireless network may result in reduced costs for telecommunication companies.**

Keywords -- cost optimization; wireless network; access point, heuristic algorithm; experimentation system.

I. INTRODUCTION

Nowadays the wireless network technology has a great impact on our life. It could carry broadband internet access to humans in the remote area network. We deal with wireless communication networks in the level of single user (PAN), local area (LAN), metropolitan area (MAN) and wide area network (WAN). In this paper, we focus on metropolitan area networks and consider the issue of optimal access point (AP) usage. The equivalent term for AP is transmitter. One of the most important parameters of the access point is its range of service. The range of an AP is related to its cost. Here, we assume that access points provide service with IEEE standards: 802.11a, 802.11b, and 802.11g. In this paper, the objective is to select and allocate the APs in a way to minimize the total cost of their distribution.

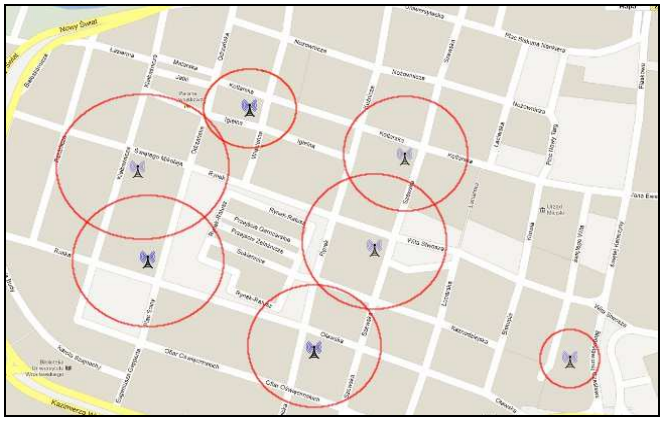


Figure 1. Example of access-points (APs) distribution.

Fig. 1 shows an example of the access point distribution that could provide network coverage inside the circles.

There are many related works in which the improvement of network design by taking into account costs of APs is discussed. Prommak and Wechtaison [1] proposed the WiMAX network design for the cost minimization and access data rate using mulit-hop relay stations. Regula and others [2] proposed algorithms for AP location, including the algorithm that uses term 'center of mass' and greedy approach. Kraimeche and Chang [3] suggested an approach for the optimization of a wireless access network based on a simple exhaustive search algorithm. Calegari and others [4] described optimisation algorithms for radio network planning, including a greedy algorithm and genetic approach. In this work, two meta-heuristic algorithms based on simulated annealing and genetic ideas are presented. The properties of these algorithms are tested with the designed and implemented experimentation system.

This paper is organized as follows. In Section II, we provide the formulation of the considered problem. The created meta-heuristic algorithms are presented in Section III. The experimentation system for testing these algorithms is described in Section IV. In Section V, the results of investigations made with this system are discussed. Section VI sums up the work and Section VII concerns further research in the area.

II. STATEMENT OF COST MINIMIZATION PROBLEM

The goal is to minimize the total cost of distribution access points over a given area. We express our objective function as:

$$Minimize\ f = min \sum_a z_a g_{ab} c_b \qquad (1)$$

where:

- a = 1, 2,…, A denotes a set of candidate APs locations,
- b = 1, 2,…, B denotes a set of APs types,
- c = 1, 2,…, C_b denotes a set of APs types' costs (for each b an access point installation cost is defined),
- g_{ab} is binary variable; it is equal to 1, if AP is installed in location a and is of b type; it is 0, otherwise,
- z_a is also binary variable; it is equal to 1, if AP is installed in location a, it is 0, otherwise.
- t = 1, 2,…, T denotes a set of TP (receiver, test point), i.e., potential users of network facilities,

 ISBN: 978-1-61208-203-5

- x_{at} is a binary variable, it is equal to 1 if TP is assigned to AP installed in location a; it is 0, otherwise
- $s = 1, 2,\ldots, S_b$ denotes a set of ranges for every b type of AP,
- $i = 1, 2,\ldots, I_a$ denotes a set of access point types assigned for each a,
- $m = 1, 2,\ldots, M_b$ denotes a set of maximum numbers of TPs possible to serve by b type AP.

Moreover, the following assumptions are taken into consideration: (*i*) the coverage of the AP is a circle (other propagation models are not considered here); (*ii*) the following constraints (2) – (7) have to be satisfied:

- The TP can be assigned only to an installed AP (i.e., $z_a = 1$) :

$$x_{at} \le z_a, \quad a = 1, 2,\ldots, A \quad t = 1, 2,\ldots, T \tag{2}$$

- It is a limit m_a for each a on maximum number of serving TP

$$\sum_t x_{at} \le m_a, \quad a = 1, 2,\ldots, A \quad t = 1, 2,\ldots, T \tag{3}$$

- Each TP can be assigned to maximum one AP:

$$\sum_a x_{at} \le 1, \quad a = 1, 2,\ldots, A \quad t = 1, 2,\ldots, T \tag{4}$$

$$x_{at} \le z_a, \quad a = 1, 2,\ldots, A \quad t = 1,2,\ldots, T \tag{5}$$

$$\sum_a x_{at} \le 1, \; a = 1, 2,\ldots, A \quad t = 1, 2,\ldots, T \tag{6}$$

$$\sum_t x_{at} \le m_a, \; a = 1, 2,\ldots, A. \tag{7}$$

III. META-HEURISTIC ALGORITHMS

Meta-heuristic approach is very often used in optimization. Formulation of the problem shows that we have to find the solution in huge discrete search-space. In this section, we present meta-heuristic algorithms to solve the problem. Firstly, we present basic procedures; next, we describe the way of adaptation of intelligent methods to creation of the algorithms.

A. Basic procedures

In the considered problem, the cost optimization is possible due to minimizing redundancy for transmitters' coverage. To achieve this goal we base on MIS (Maximum Independent Set) model used in [4]. The classical scheme for MIS extraction can be found in [5]. The concept of solving problem considered in this paper is lying in creating possible independent sets of receivers and then matching to them the cheapest possible transmitters. Firstly, for all receivers has to be created neighbourhood table (NT). NT consists of neighbourhood lists (NL). Each receiver has its own NL. Such list stores all neighbours of particular receiver. Neighbourhood is defined on the basis of largest range from set *s*. It means that for a given receiver all TP in distance less than S_{max} are its neighbours.

The procedure of creating complete neighbourhood table (CNT) for all receivers in *t* can be described as follows:

1. Find a largest range S_{max} in set *s*.
2. For each receiver from set *t* check whether distance between a given receiver T_t and the rest of TPs in set *t* is less than S_{max}.
3. If yes, add indices of these receivers to list of neighbours for a given receiver T_t.

The procedure of creating releasing neighbourhood table (RNT) can be described as follows:

1. Choose NL from NT for processing.
2. Match the cheapest transmitter that can support all receivers on the list.
3. Delete receivers from processed NL (do it for whole NT)
4. If NT is not empty go to step 1

As the result of this approach, an independent sets of receivers can be acquired. Next, AP is placed in the candidate AP location nearest the centre of set of receivers given in NL. Note that the chosen transmitters (APs) differ in dependence of order of processing the NL. The cost is a sum of installation costs *c* for types of AP selected in RNT.

The sequence of mentioned here procedures ends with calculated final cost and it is taken as a criterion function. In contrast to other AP distribution optimization methods, we do not take as a search space set of candidate AP locations but processing order for NL. And this sequence is a discrete search-space searched for the optimal solution in two implemented meta-heuristic algorithms.

B. Algorithm based on Simulated Annealing (SA)

The idea of simulated annealing method was used because of simplicity and efficiency [5]. In this method, as the criterion cost function the system energy is interpreted. The system state can be understood as the processing order of NL internal parameters of SA. In our implementation, internal parameters of SA were discovered experimentally. Entering perturbation (EP) is done in the following way:

1. Choose two random NL in NT
2. Reverse order between chosen NLs

The presented form of EP allows eliminating weakness of classic simple replacement of two chosen NL. Such classic replacement does not cause big change in search space.

C. Algorithm based on Genetic Algorithms

We applied the population-based model [6] that uses selection and recombination operators to generate new feasible solution. The roulette wheel selection is used to evaluate the fitness value associated with each individual (chromosome): the

 ISBN: 978-1-61208-203-5

higher the fitness value of an individual, more likely it is to be selected. In our approach, chromosomes can be defined as processing order for neighbours' lists. Crossover is a combination of two chosen individuals and mutation is a random disturbance in this individual. The program execution stops when the predefined number of iteration steps has been run through [4]. Outline of the algorithm is presented as below.

1. Choosing random population of *n* chromosomes (feasible solutions for the problem).
2. Evaluate fitness (cost function computed) of each individual chromosome *x* in the population.
3. Repeat the steps below until the new population is created.
   3.1. Select pairs of chromosomes from a population according to their fitness ranking.
   3.2. According to crossover probability cross over pairs of chromosomes to form a new offspring.
   3.3. Apply mutation operation within a new offspring.
   3.4. Put new offspring in a new population.
4. Replace an old population by the new generated one.
5. Check whether the stop condition is satisfied. If yes, then stop. Return the best solution.
6. Go to step 2.

IV. EXPERIMENTATION SYSTEM

The model of the experimentation system is shown in Fig. 2. The idea of the system proposed in [7] was applied in order to allow testing properties of the created algorithms.

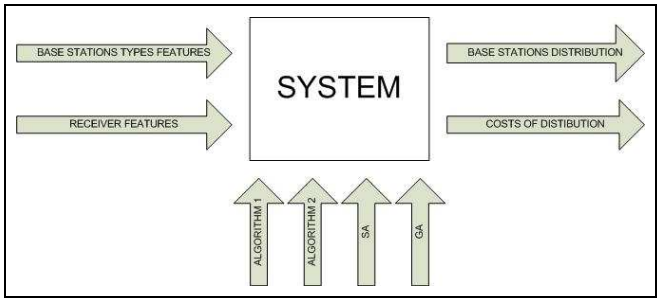


Figure 2. Experimentation system.

Input parameters of the system (simulator) are:

- AP types features
- Receivers features

Output parameters of the system (simulator) are:

- AP distribution
- Cost of distribution

The simulator was implemented in .NET C# in Visual Studio 2008 IDE. Simulator has graphical user interface (see Fig. 3), which allows choosing configuration options (design of experiment). Algorithms were tested under Microsoft Windows 64-bit OS, CPU of 1 GHz and 2 GB of RAM. Recently, the implemented simulator gives possibilities for observing effects of using four algorithms to finding location of APs (denoted as base stations in Fig. 2), including two created algorithms:

- Algorithm based on Simulated Annealing (SA),
- Algorithm based on Genetic Ideas (GA),

and two algorithms described in [2] [7] :

- Algorithm 1 (Alg. 1) - It is based on the physic's term – center of mass. The set of possible receivers is divided into the subsets. The subset's center of mass is considered as the location of the AP. Details are given in [7].
- Algorithm 2 (Alg. 2) - It is a greedy algorithm, which uses a particular graph model in order to implement adjacency matrix. This matrix is being checked in order to find the cheapest AP that supports receivers in nearby area [7].

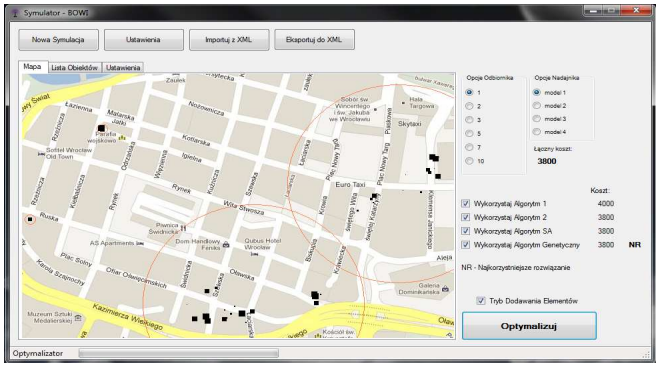


Figure 3. Simulator interface - working environment.

V. INVESTIGATIONS

The aim of the investigations was comparing efficiency of different methods for solving problem of location APs (access point's distribution) concerning minimization of the cost defined by (1) with assumptions expressed by (2) – (7).

Due to the fact that in real life receivers are distributed in many different ways, we provided differentiated benchmarks. To realize this task we created and implemented a module of experimentation system for generating different types of distribution. It allows testing algorithms in various conditions. Benchmark generator ensures 4 types of distribution. The user of experimentation system can set parameters such as the total number of receivers and the introduced distribution type:

- Regular – receivers are distributed regularly on the map, distances between points on the map are equal and in each point the total number of receivers is equal too;
- Regular with randomness – philosophy of this distribution is almost the same like in the regular distribution, but numbers of receivers differ – they are chosen at random;
- Irregular – locations of receivers are chosen randomly, but constraints arising from set parameters are held;
- Grouped – locations of receivers are grouped; groups are confined by surface of random circle shape.

 ISBN: 978-1-61208-203-5

A. Costs

For each series of experiments the total number of receivers was different – in range from 100 to 1000 with step 100. The obtained results are shown in Fig 4, Fig. 5, Fig. 6, and Fig. 7.

Experiment 1 – Regular distribution

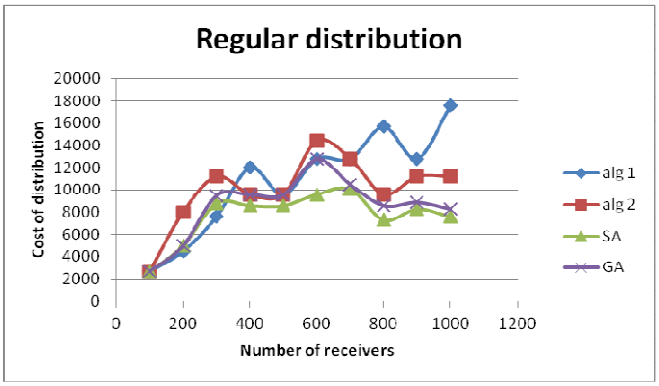


Figure 4. Regular distribution - cost in relation to the number of receivers.

Experiment 2 – Regular distribution with randomness

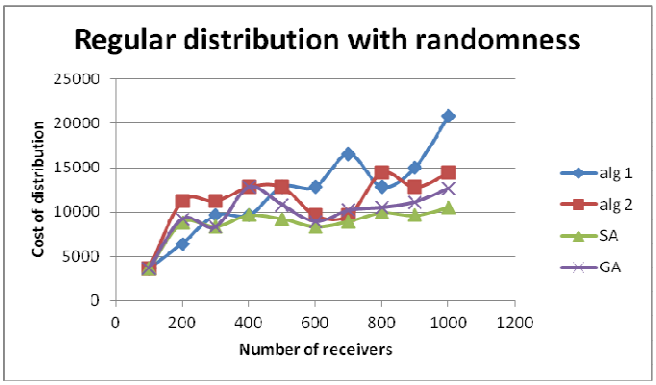


Figure 5. Regular distribution with randomness - cost in relation to the number of receivers.

Experiment 3 – Irregular distribution

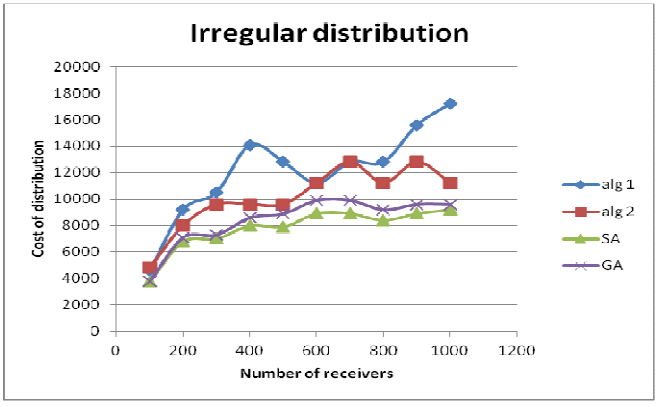


Figure 6. Irregular distribution - cost in relation to the number of receivers.

Experiment 4 - Grouped distribution

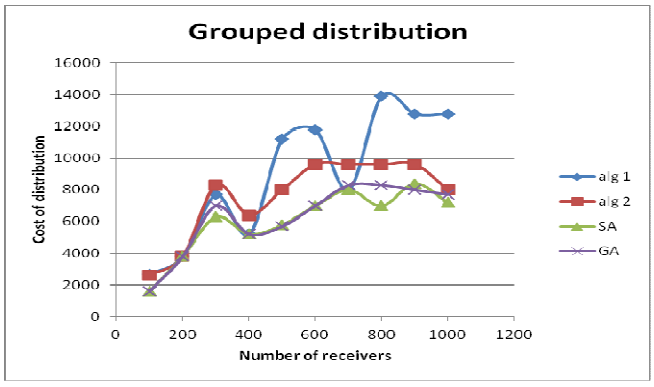


Figure 7. Grouped distribution - cost in relation to the number of receivers.

It may be observed that almost for every tested distribution type, the newly implemented algorithms decreased distribution cost. It is worth to be noticed that SA performed a little bit better than GA, especially in the case of irregular distribution – the most common distribution in the real world. Observing the shapes of graphs for SA and GA, one may conclude that they are quite similar to those 'produced' by Algorithm 2.

B. Execution time

The execution time for the both developed heuristic algorithms significantly increased in comparison with heuristic Algorithm 1 and Algorithm 2. The scale of this increase is visible in Fig. 8 and in Table I. Using meta-heuristic algorithms requires about 100 times longer execution time (in average). Notwithstanding to this fact, the execution time is still short – the longest was of about 1 sec. However, the obtained cost profits are worth this increment. One should remember that this paper deals with optimization of base station distribution cost – making decision about location of APs (distribution of base stations) is long-term designing process and it is worth to devote more computation time to obtain better solution to the considered problem.

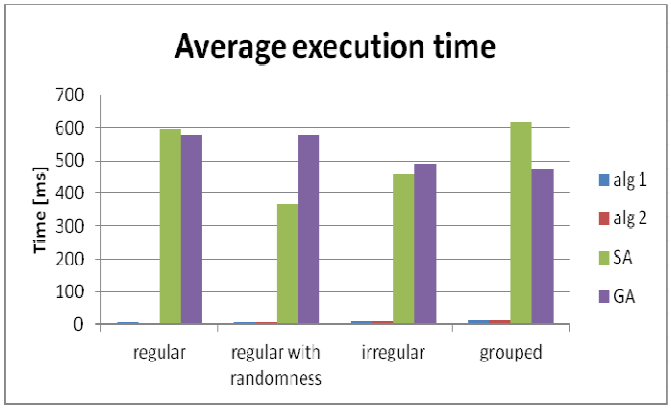


Figure 8. Comparison of the averaged execution times for different distributions.

 ISBN: 978-1-61208-203-5

TABLE I. AVERAGE EXECUTION TIME

The averaged execution time [ms]					
		Alg 1	**Alg 2**	**SA**	**GA**
Distribution type	**Regular**	5.75	3.22	594.9	578.5
	Regular with randomness	6.85	5.61	369.0	580.2
	Irregular	10.37	10.60	460.3	490.6
	Grouped	12.17	13.58	617.2	473.3

C. Improvement

To show the obtained improvement while using the proposed approach (meta-heuristic algorithms), we present the results of comparison between algorithms.

Comparison to Algorithm 1. In comparison with Algorithm 1, the remarkable dependency between improvement and the number of receivers can be observed (see Fig. 9 and Fig. 10). This relationship can be expressed as follows: the larger number of receivers, the bigger improvement is obtained. It goes up to 60%. Such improvement can bring huge financial benefit.

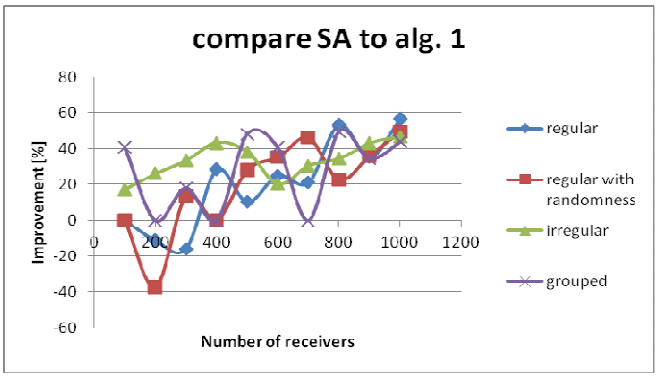


Figure 9. Improvement for SA and Alg. 1.

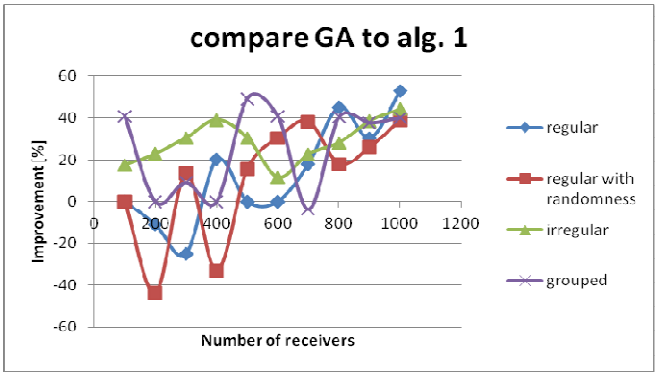


Figure 10. Improvement for GA and Alg.1.

Comparison to Algorithm 2. In comparison to Algorithm 2, such clearly relation like in comparison to Algorithm 1 is not visible (see Fig. 11 and Fig. 12). The improvement is unstable and does not depend on the number of receivers but in most of cases is positive and up to 40%.

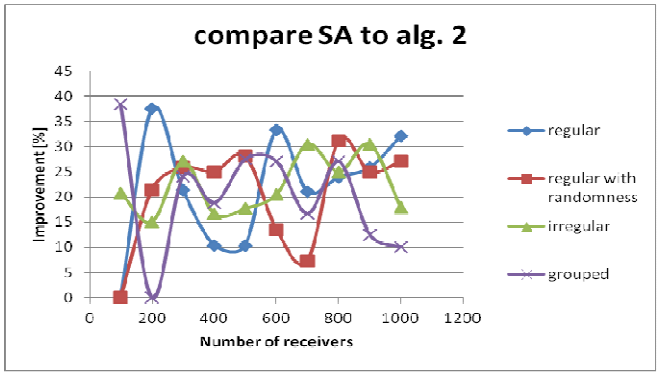


Figure 11. Improvement for SA and Alg. 2.

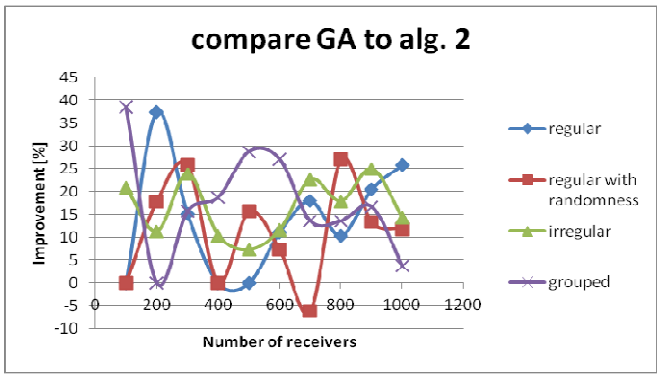


Figure 12. Improvement for GA and Alg. 2.

In Table II, the comparison of general (over all distribution types) improvements is presented. In columns, the created meta-heuristic algorithms applied to the problem and in rows, previously used Algorithm 1 and Algorithm 2, are specified. In the intersections, are improvement percentages between them, e.g., in the intersection between SA and Algorithm 1 is result 25%, it means that SA improved average distribution cost in comparison to Algorithm 1 by 25%.

TABLE II. COMPARISON BETWEEN THE IMPLEMENTED ALGORITHMS

	General average improvement of algorithms	
	SA	**GA**
Algorithm 1	25%	19%
Algorithm 2	21%	15%

Observing results shown in Table II, we may conclude that the average improvement may be estimated as around 20%.

 ISBN: 978-1-61208-203-5

VI. CONCLUSION

The aims of conducted investigations - presented in this paper - were achieved. We shown the advantage of the proposed approach – using of the created meta-heuristic algorithms ensured that the average improvement was of around 20%. It may be interpreted as a big advancement. The obtained results show that use of meta-heuristic algorithms in presented problem is profitable. Rising the execution time is the fact, but it can not be not acceptable when does not exceed 1 sec.

The weak points of investigation presented in this paper are lying in: taking into consideration a simple model and focusing only on cost optimization; algorithms were tested only on artificial generated benchmarks (basing on simulation experiments) – it would be more reliable to use real data.

VII. FUTURE WORK

The experimentation system can be developed in many ways as an incremental approach to aiding optimization of AP location problem. The system can be improved by including modules giving possibilities of aiding process of solving more complicated problem with propagation model applied [8] and quality indicators taken into consideration.

We believe also that results obtained by heuristics algorithms could be improved in some cases by two stage approach [9], including adjustment of internal parameters [10] of the meta-heuristic algorithms at the first stage. The efficiency of the Algorithm based on Simulated Annealing, largely depends on three internal parameters: cooling factor, start temperature and end temperature. The similar situation concerns the Algorithm based on Genetic Algorithm. Thus, we plan making more experiments with changing the number of iterations in which this algorithm creates new population and also more experiments which results could give hints how to match the size of chromosome's population. Finding optimal internal (input) parameters for both algorithms could be time-consuming; however, it may result in further minimization of the cost of the designed location of the base stations.

Moreover, some improvements in application are planned, e.g., the input map of simulator could be downloaded through Google Maps Api. It would improve the usability of graphical interface; maps would be created dynamically, either. It is also desirable to have such possibilities as saving all experimentation results in data base and generating charts in 'automatic way'.

The experimentation system with the simulator as a core is an aiding tool in teaching process at the Faculty of Electronics, Wroclaw University of Technology.

ACKNOWLEDGMENT

The authors of this paper would like to thank Agnieszka Beza, Michal Michalski and Krzysztof Pajak - students of the Faculty of Electronics, Wroclaw University of Technology, and authors of simulator [7], for availability to use their application for further development.

REFERENCES

[1] Ch. Prommak and Ch. Wechtaison, "WiMAX Network Design for Cost Minimization and Access Data Rate Guarantee Using Multi-hop Relay Stations", International Journal of Communications, vol. 4, 2010.

[2] P. Regula, I. Pozniak-Koszalka, L. Koszalka, and A. Kasprzak, „Evolutionary Algorithms for Base Stations Placement in Mobile Networks", Intelligent Information and Data Base Systems, Springer, Lecture Notes in Computer Science, vol. 6592, pp. 1-10, 2011.

[3] K. Kraimeche, B. Kraimeche, and K. Chiang, "Optimization of a Wireless Access Network", Proceedings of IEEE Systems and Engineering Design Symposium, James Madison University, Virginia, USA, April, 2006.

[4] P. Calégari, F. Guidec, P. Kuonen, B. Chamaret, S. Josselin, D. Wagner, and M. Pizarosso, "Radio Network Planning with Combinatorial Optimization Algorithms", Proceedings of the 1st ACTS Mobile Telecommunications Summit, Chr. Christensen, Ed., vol. 2, pp. 707-713, November, 1996.

[5] T. Spencer and M. Goldberg,, "A New Parallel Algorithm for the Maximal Independent Set Problem", SIAM Journal, vol. 54, pp. 11-121, 1989.

[6] H. Cormen, C.E. Leiserson, and R.L. Rivest, "Introduction to algorithms", MIT Press, 1990.

[7] A. Beza, M. Michalski, and K. Pajak, "Application to Simulate Wireless Network in Regard to Cost optimization for Base Stations Distribution.", Report of Dept. of Systems and Computer Network, Wroclaw University of Technology, Wroclaw, 2009.

[8] E. Osekowska, I. Pozniak-Koszalka, and A. Kasprzak, „Impact of Propagation Factor on Routing Efficiency in Wireless Mesh Networks", Proceedings of IARIA. 11th International Conference on Networking (ICN'12), 2012.

[9] L. Koszalka, D. Lisowski, and I. Pozniak-Koszalka, „Comparison of Allocation Algorithms for Mesh Networks with Multistage Experiments", Springer, Lecture Notes in Computer Science, vol. 3984, pp. 58-67, 2006.

[10] L. Koszalka, M. Kubiak, and I. Pozniak-Koszalka, "Allocation Algorithm for Mesh-Structured Networks", Proceedings of IARIA 5th International Conference on Networking, IEEE Comp. Society Press, pp. 24-29, 2006.

 ISBN: 978-1-61208-203-5

Multi-Level Collaborative Spectrum Sensing in Nakagami Fading Channels

Omkalthoum El-Bashir Hamed
Department of Electrical Engineering
The Petroleum Institute
Abu-Dhabi, United Arab Emirates
e-mail: ohamed@pi.ac.ae

Mohammed Abdel-Hafez
Department of Electrical Engineering
United Arab Emirates University
Al-Ain, United Arab Emirates
e-mail: mhafez@uaeu.ac.ae

***Abstract*— This paper is to investigate the problem of spectrum scarcity and underutilization with particular attention to the performance of opportunistic spectrum access in fading channels. In this paper we studied the energy detector in collaborative and non-collaborative sensing modes when the channels between the primary and the sensors are generalized Nakagami-*m* fading channel. Soft combining techniques perform well enough, but require that each spectrum sensor sends complete signal information to the band manager. Sending the signal information introduces unnecessary complexity. Moreover, it is more complicated and time consuming for the band manager to handle. To reduce the communication overhead, hard decision techniques can be used. In this paper, two techniques will be studied. The first is the simple hard decision technique, and the second is the use of multi-threshold decision technique. The results of this study show that the multi-threshold technique outperforms the single one with slight increment in the cost. The performances of all these techniques are evaluated in terms of probability of false alarm and probability of detection. Although soft decision techniques give less probability of miss detection at certain value of probability of false alarm, the hard techniques are simpler to implement. It is also found that the multi-threshold works better than the single threshold especially in low SNRs.**

Keywords - Spectrum sensing; opportunistic access; cognitive radio; Nakagami-m fading channel; square law combining; maximum selection combining; hard decision combining.

I. INTRODUCTION

The underutilization of the spectrum leads to thinking in managing the spectrum in more flexible way by allowing second level of spectrum usage. So, some users called "secondary users" are allowed to access spectrum holes, a band of frequency assigned to a primary user, but at a particular time and specific geographic location, the band is not being utilized by that user [1, 2]. Spectrum utilization can be improved significantly by making it possible for secondary users to access spectrum holes.

The telecommunication sector debates the reallocation of frequencies used for GSM, plans for digital TV switchover, formulates policies for cognitive radio and considers options for dealing with the wireless data explosion [3]. So, applying a sensing technique in the opportunistic wireless networks is needed. A simple sensing technique that can be used for this purpose is the energy detection. One of the simplest energy detectors is presented in [4,5]. This energy detector measures the energies of $2N$ samples of a received signal over a flat band-limited Gaussian noise channel. Then, it combines these Gaussian samples for comparison with a certain threshold. The result of the comparison can be defined by two hypotheses, either signal or no signal. Accordingly, the secondary user will decide on whether or not to access the spectrum band. Relying on chi-square statistics of the resulting sum of squared Gaussian random variables, Urkowitz [6] derived both probability of detection and probability of false alarm in Gaussian channel. Since we need to have enough protection for the primary user, we have to set the threshold such that it provides some protection level to the primary user. However it is found that one sensor cannot provide reliable sensing system specially in real fading channels. Collaborative sensing techniques had been studied in [8,9,10] with local energy detectors to improve the sensing system performance in fading channels. Zou et al. [8] investigated the effect of user collaboration in Rayleigh fading channel. It showed that using more collaborative users increases the performance significantly and improves the spectrum utilization. The researcher used the equal gain and maximum selection as soft decision combining techniques to combine the local measurements and finalize the decision. Although, soft combining techniques perform well, it consumes high bandwidth for sharing information. Simple hard decision combining technique was used by [8] where each user shares his vote with the controller using only binary 0 for empty and binary 1 for occupied band. Then, the controller makes the decision according to the collaborative users votes. As a comparison between the soft and hard decision combining techniques, soft techniques give much better decisions, but the cost is in the network overload by the overhead used to share their knowledge about the signal to noise ratios.

Yilmaz et al. [7] extended the work by applying a new combining technique which is collaborative sensing with a decision vector that uses a uniform quantization. It consists of multiple thresholds and a weight vector for global decision. Each operating secondary user should sense the channel locally and decide on one of the designed levels according to the measured signal to noise ratio. Then the secondary user should send his decision to the fusion center. The fusion center then makes final decisions according to the different users' votes using a special weighted sum decision rule. This method performed better than the single threshold hard decision studied in [8] with little addition to the overhead. Moreover, as much as the number of levels increased, the decision becomes better and the overhead increases.

In [9], Liang et al. came up with a closed form solution for the probability of detection in Nakagami-m channel with

 ISBN: 978-1-61208-203-5

only integer fading parameters in non-collaborative mode. Since Nakagami-m fading channel can have a non-integer fading parameter, and to our best knowledge, this hasn't been studied this before. Therefore we started the work by using new approach to study soft combining collaboration techniques in sensing in Nakagami-m fading channel with any real parameter. In this paper, we extended the work of [11] to study the hard single and multi-thresholds combining techniques for collaborative sensing in Nakagami fading channel with any real fading parameters.

The rest of this paper is organized as follows: Section 2 presents the system model followed by the analysis of spectrum sensing in Additive White Gaussian Noise (AWGN) channel in Section 3. Section 4 addresses the spectrum sensing in Nakagami fading channel. Single threshold hard decision detection technique is introduced in Section 5. In Section 6, multi-levels hard decision technique is introduced and the effect of collaborative sensing is studied. Section 7 shows the results of some numerical examples. Concluding remarks are presented in Section 8.

II. System Model

Figure 1 shows the suggested model for sensing scenario in our opportunistic spectrum access system. Infrastructure-based sensors are distributed in the model to sense the primary user signal in a certain band. The channel between the sensor and the use is assumed to be generalized Nakagami-m fading channel with instantaneous signal to noise ratio γ. The band manager at data fusion center will then decide based on the information reported from k sensors with one of combining techniques will be studied in this paper.

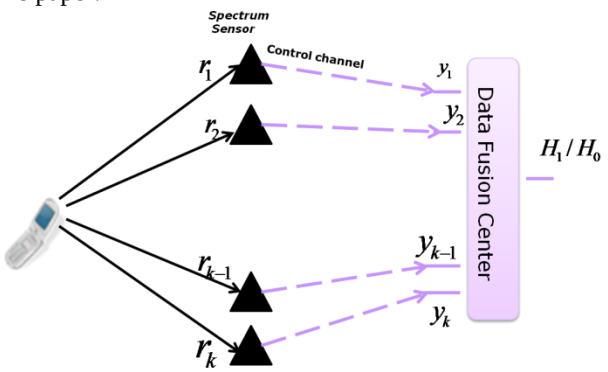


Figure 1. System model.

III. Spectrum Sensing in Non-Collaborative Mode in AWGN

AWGN channel is the ideal case of wireless channel where the noise is only due to the additive noise at the receiver. The performance in AWGN channel is studied in terms of the probability of detection and probability of false alarm. The probability of detection, P_d, and probability of false alarm, P_f, in AWGN channel were studied in [6]. Then it is revisited by [9] and studied for the sampled version of the signal. It is found that these probabilities can be expressed as,

$$P_d = Q_N\left(\sqrt{\frac{2\gamma}{\sigma^2}}, \sqrt{\frac{\lambda}{\sigma^2}}\right) \quad (1)$$

and

$$P_f = \frac{\Gamma(N, \lambda/2\sigma^2)}{\Gamma(N)} \triangleq G_N(\lambda) \quad (2)$$

where γ is the SNR, σ^2 is the variance of the channel. For simplicity and without loss of generality, σ^2can be assumed to be unity, $\Gamma(.,.)$ is incomplete gamma function [12], N is the half number of samples. $Q_N(.,.)$ is the Generalized Marcum Q function [13].

IV. Spectrum Sensing in Nakagami Fading Channels

This section focuses on the performance of spectrum sensing in the Nakagami fading channel. The performance of spectrum sensing in the Rayleigh fading channel is a special case when fading parameter $m = 1$. The performance is formulated in terms of the probability of detection and probability of false alarm.

For the Nakagami fading channel, the pdf of the signal to noise ratio SNR, γ, has the following gamma distribution,

$$f_\gamma(\gamma) = \frac{1}{\Gamma(m)}\left(\frac{m}{\bar{\gamma}}\right)^m \gamma^{m-1}\exp\left(\frac{-m}{\bar{\gamma}}\gamma\right), \gamma \geq 0 \quad (3)$$

where $\bar{\gamma}$ is the average signal to noise power ratio in the fading channel and m is the Nakagami fading parameter.

The probability of false alarm is independent of γ because it is the probability of the received energy being above the threshold with the absence of the primary user. Since, under H_0, no primary user's signal exists, P_f is not affected by fading. On the other hand, the probability of detection over Nakagami fading channel, $P_{d,Nak}$, can be found by averaging (1) over (3) as,

$$P_{d,Nak} = \frac{1}{\Gamma(m)}\int_0^\infty Q_N\left(\sqrt{(2\gamma)}, \sqrt{\lambda}\right)\left(\frac{m}{\bar{\gamma}}\right)^m . \gamma^{m-1}\exp\left(\frac{-m}{\bar{\gamma}}\gamma\right) d\gamma \quad (4)$$

This integration can be simplified using the following steps: Let $x = \sqrt{(2\gamma)}$, $b = \sqrt{\lambda}$, and $p^2 = \frac{m}{\bar{\gamma}}$, then (4) can be expressed as,

$$P_{d,Nak} = \frac{2}{\Gamma(m)}\left(\frac{p^2}{2}\right)^m \int_0^\infty Q_N(x, b)\, x^{2m-1} . \exp\left(-\frac{p^2x^2}{2}\right) dx \quad (5)$$

Marcum Q-function defined by [12] is used and defined by,

$$Q_N(x,b) = \int_b^\infty \frac{\alpha^N}{x^{N-1}} e^{-\left(\frac{x^2+\alpha^2}{2}\right)} I_{N-1}(\alpha x)\, d\alpha \qquad (6)$$

where α is a dummy variable. After some manipulations and using the formulas in (p720 and p1059, [15]), the probability of detection in Nakagami channel can be expressed as,

$$P_{d,Nak} = \frac{2^{-N+1}}{\Gamma(N)}\left(\frac{p^2}{1+p^2}\right)^m . \int_b^\infty \alpha^{2N-1} e^{-\frac{\alpha^2}{2}} \varphi\left(m, N, \frac{\alpha^2}{2(1+p^2)}\right) d\alpha \qquad (7)$$

where $\varphi(.,.,)$ is the degenerate hyper geometric function defined in [12].

The probability of miss detection, can be expressed as,

$$P_{m,Nak} = 1 - P_{d,Nak} \qquad (8)$$

or

$$P_{m,Nak} = \frac{2^{-N+1}}{\Gamma(N)}\left(\frac{p^2}{1+p^2}\right)^m . \int_0^b \alpha^{2N-1} e^{-\frac{\alpha^2}{2}} \varphi\left(m, N; \frac{\alpha^2}{2(1+p^2)}\right) d\alpha \qquad (9)$$

reverting to the original terms and constants in (9), we arrive at the following,

$$P_{m,Nak} = \frac{2^{-N+1}}{\Gamma(N)}\left(\frac{m}{\bar{\gamma}+m}\right)^m . \int_0^{\sqrt{\lambda}} \alpha^{2N-1} e^{-\frac{\alpha^2}{2}} \varphi\left(m, N; \frac{\alpha^2\bar{\gamma}}{2(\bar{\gamma}+m)}\right) d\alpha \qquad (10)$$

The integration in (10) is limited and can be evaluated easily by using the Monte Carlo integration method [16]. The advantage here is that $P_{m,Nak}$ can be evaluated for (integer and non-integer) fading parameters m. At this point, the probability of miss detection in Rayleigh fading channel can be found by simply setting $m = 1$ as special case. This approach of finding the probability of miss detection in Nakagami-*m* fading channel is suggested and confirmed by comparing its results with some results in the literature. Digham and Alouini [4] found a closed form formula for the average probability of detection in Nakagami for only integer values of m values.

V. SINGLE THRESHOLD HARD DECISION TECHNIQUE

Using this technique, the spectrum sensor sends only one bit information as an individual decision. It sends 0 if the locally detected signal energy is less than the threshold to decide on H_0. Otherwise, it sends 1 to decide on H_1. Then, the band manager finalizes the decision using votes according to the "n out of k" rule, where n is the required number of voters necessary to decide on the existence of the primary signal.

Given that all the sensors are independent, and applying the Neyman-Pearson criterion (which is based on fixing the probability of false alarm to an acceptable value to find a test threshold that maximizes the probability of detection), results in the following combining rule [5]

$$\sum_{i=1}^{n} S_i \log_e\left[\frac{P_{d_i}(1-P_{f_i})}{(1-P_{d_i})P_{f_i}}\right] \underset{H_0}{\overset{H_1}{\lessgtr}} \Lambda \qquad (11)$$

where S_i is the i^{th} sensor decision. P_{d_i} and P_{f_i} are the individual probabilities of detections and false alarm, respectively. The band manager decides by comparing the weighted sum of the individual decisions to a threshold Λ; where Λ is a global threshold with discrete value in this technique. In this study, iid sensors are assumed to simplify the analysis. So, P_{d_i} and P_{f_i} are assumed to be equal for all the sensors as a result of identical path loss and fading. It is also assumed that all the users imply the same threshold λ in their local decision for simple implementation. Thus based on the other chosen global threshold Λ, the data fusion center implements an "n out of k" voting rule. It decides H_1 if n or more vote to H_1, otherwise, it will decide on H_0. The average probabilities of detection and false alarm for the n out of k rule are related to their single user probabilities through binomial distribution. The AND and the OR decision rules are considered as special cases from the general n out of k rule. By using AND rule, the band manager will decide on H_1 if all the sensors agree on deciding on the primary user existence. On the other hand, by the OR rule, the band manager will decide on H_1 when at least one sensor has decided locally on the primary user existence [5]:

$$P_{f_{Nak,HD}} = \sum_{i=n}^{k} \binom{k}{i} {P_f}^i \left(1-P_f\right)^{\mathrm{k-i}} \qquad (12)$$

and

$$P_{d_{Nak,HD}} = \sum_{i=n}^{k} \binom{k}{i} {P_{d,nak}}^i \left(1-P_{d,nak}\right)^{\mathrm{k-i}} \qquad (13)$$

where $P_{d,Nak}$ and P_f are the individual probabilities of detection and false alarm as defined by (7) and (2), respectively.

VI. MULTI-THRESHOLD HARD DECISION TECHNIQUE

The simple hard decision algorithm introduced in Section 5 was based on single threshold. In the single threshold detection method, the decision H_0 or H_1 depends only on one local threshold λ. P_d and P_f for a single secondary user

 ISBN: 978-1-61208-203-5

can be calculated using λ for a selected channel model and using the exact formulas given by (12) and (13).

Figure 2 shows four-thresholds as an example of the method proposed by [7], namely, λ_0 , λ_1 , λ_2 and λ_3. The distance between the center thresholds λ_1 and λ_2 and the other thresholds is fixed and is equal to Δ, and there is a Δ_c distance among the center thresholds themselves. Each sensor determines the quantization bin from the bins vector $[B_0, B_1, B_2, B_3, B_4]$ locally according to its measurement and the thresholds given. For example, if the measured energy value is between the values of λ_0 and λ_1 , the sensor will decide on B_1 bin. When the measurement is in the region between thresholds λ_1 and λ_2 the sensor shouldn't send its decision. This technique censors some sensors from sending their information because of its low importance. The measurements in the "no decision" region are not important because it is in the middle of the range. In other words, it is not high enough to vote for the primary user presence, nor low enough to vote for its absence. The idea behind censoring is avoiding overloading the band with unnecessary data by having sensors send their decision to the fusion center only if this decision is considered to be "informative" Or, only if they are sure enough about it.

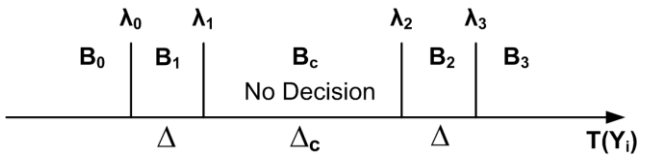


Figure 2. Multi threshold Energy detector with four thresholds.

The sensors send their softened decisions or quantized measurements in 2-bit formats for 4-threshold case and 3-bit formats for 8-threshold cases and so on. The fusion center gives a weight vector to the quantization bins. So, deciding on each bin has special weighting determined by the fusion center to change the rule of the decision used. For example $\vec{w} = [-1\ -1\ 0\ 1\ 1]$ is equivalent to the majority rule. The fusion center receives the softened decisions and counts the number of users in each quantization bin and forms a vector $\vec{B}$ that lists how many sensors reported in each bin. Then, if the inner product of the two vectors $\vec{w}$ and $\vec{B}$ is > 0, $\delta_{\vec{w}}(\vec{B})$ is considered 1, otherwise, it will be considered as 0.

To quantify the performance of this method, probability of detection and probability of false alarm are calculated using the formulas below [7]:

$$P_{d=}\sum_{\vec{B}\in\beta}\delta_{\vec{w}}(\vec{B})\,C(k,n_0)C(k-n_0,n_1)C(k-n_1-n_0,n_2)$$
$$C(k-n_2-n_1-n_0,n_3)\left(P_{B_0,H_1}\right)^{n_0}\left(P_{B_1,H_1}\right)^{n_1}\left(P_{B_c,H_1}\right)^{n_c}$$
$$\left(P_{B_2,H_1}\right)^{n_2}\left(P_{B_3,H_1}\right)^{n_3} \quad (14)$$

and

$$P_f=\sum_{\vec{B}\in\beta}\delta_{\vec{w}}(\vec{B})\,C(k,n_0)C(k-n_0,n_1)C(k-n_1-n_0,n_2)$$
$$C(k-n_2-n_1-n_0,n_3)\left(P_{B_0,H_0}\right)^{n_0}\left(P_{B_1,H_0}\right)^{n_1}\left(P_{B_c,H_0}\right)^{n_c}$$
$$\left(P_{B_2,H_0}\right)^{n_2}\left(P_{B_3,H_0}\right)^{n_3} \quad (15)$$

where β represents all combinations of number of users distributed in quantization bins, $C(k;\ n)$ represents n combinations out of k, n_i represents the number of users in B_i , and P_{B_i,H_j} represents the probability of the received energy being in B_i conditioned on H_j and under AWGN channel. Similar formulas can be obtained for fading channels by calculating local probabilities according to fading channel formulas. This model is used in [7] to evaluate ROC in AWGN and in Rayleigh channels. In this thesis, only a special case of this model is studied and applied as per the Nakagami-*m* channel with any real fading parameter.

VII. NUMERICAL RESULTS

Figure 3 shows the Complementary Region Of Convergence (CROC) for combined *iid k* spectrum sensors in Nakagami fading channel with average $SNR = 20\ dB$ and fading parameter $m = 1.8$. The figure shows the performance improvement when using more than one sensor to detect the channel when the hard decision combining technique is used. The decision rule is based on "n out of k" rule. Figure 3 shows the OR rule has the least probability of miss detection at a certain value of probability of false alarm among all the other values of n. The AND rule has the most probability of detection at a certain value of probability of false alarm among the other voting rules. This means that by using the OR rule, we can guarantee a better level of QoS for the primary user. However, the down side is degradation in utilization. Therefore, the AND rule gives better utilization, but it decreases the QoS of the primary user. Performance of all other "n out of k" schemes is in between the two extreme cases, the OR and the AND rules. The network designer should be aware of the required level of primary user QoS and the additional utilization required in deciding which rule to use.

By comparing this technique with the soft decision technique presented in [11], we can consider the performance of the least probability of miss detection case (OR curve) with the soft decision cases when the number of collaborated users is $k = 4$. In general, soft decision technique outperforms hard decision but hard decision technique is much simpler. The complexity of the soft decision combining techniques arises from different factors. The first is implementation needs; for example MSC needs channel gain estimation. Moreover, in the soft decision combining technique, the process of gathering information from sensors is complicated. In addition, the sensors share their measurements rather than their decisions with the fusion center. This needs more bandwidth to carry all the

 ISBN: 978-1-61208-203-5

information, especially when there is large number of sensors. This will affect the utilization of the scarce spectrum, because the hard decision uses only one bit (0 or 1) to report its final decision to the band manager, after which the band manager simply applies one of the discussed voting rules to finalize the decision. So, a very simple receiver can carry out the process of finalizing the decision.

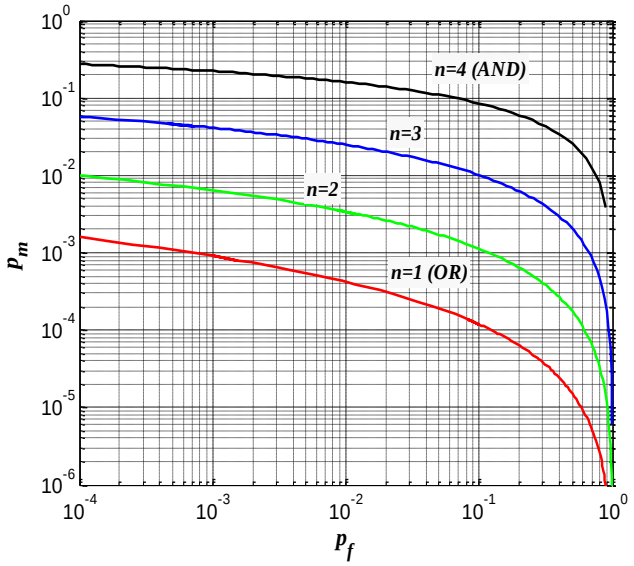


Figure 3. CROC for diversity in Nakagami for $k = 4$ iid sensors using Hard Decision combining technique for average $SNR = 20dB, N = 5$, and $m = 1.8$.

Figure 4 shows the effect of the number of collaborated sensors on detection performance. In this figure, $k = 8$ iid sensors are used. The figure shows "n out of k" voting rule with especial cases OR, and AND. The improvement in the performance due to larger number of collaborative sensors is very clear in this instance. So, to get good detection without degrading the utilization, the designer can use more sensors and use any of the n out of k. For example, instead of using the 4 out of 4 rule, we can use the 4 out of 8 rule to get better performance without degrading the utilization. The cost of this improvement is the cost of the extra sensors and a slight addition in band consumption (1 bit/sensor).

Figure 5 shows the CROC curve for the single and 4-threshold Hard Decision combining technique model system. The figure is generated for the two cases with $k = 5$ collaborative sensors, average $SNR = 20\ dB$, and majority rule. The 4-threshold majority rule is chosen with a weighting vector $\vec{w}$ =[-1 -1 0 1 1]. For the single threshold, the majority rule is considered when $n = 3$ in the "n out of k" rule and is compared with the performance of the single threshold hard decision combining technique. It is found that this method significantly outperforms the single threshold hard decision technique. The cost of this improvement is only a slight increment in the overhead coming from sending two bits instead of one bit for each sensor.

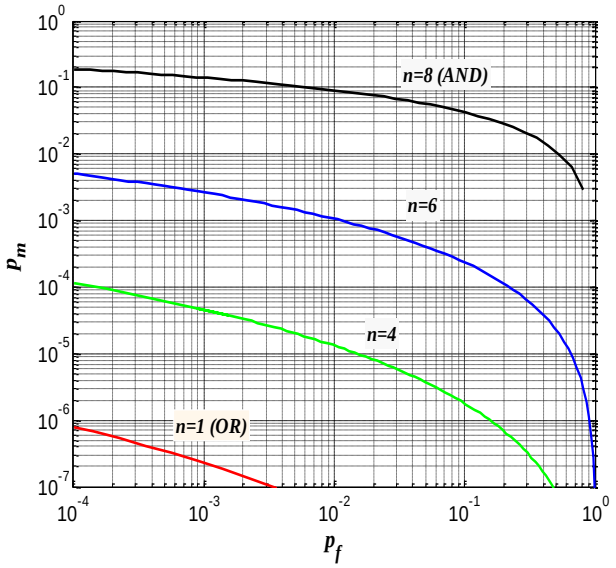


Figure 4. CROC for diversity in Nakagami for $k = 8$ iid sensors using Hard Decision combining technique for average $SNR = 20dB, N = 5$, and $m = 1.8$.

This combining technique can be considered to be the best among all the combining techniques studied, because it gives the designer the chance to make a trade-off between the cost and the detection accuracy of the system. Then accordingly, the number of collaborative sensors and number of thresholds can be chosen to fit the need.

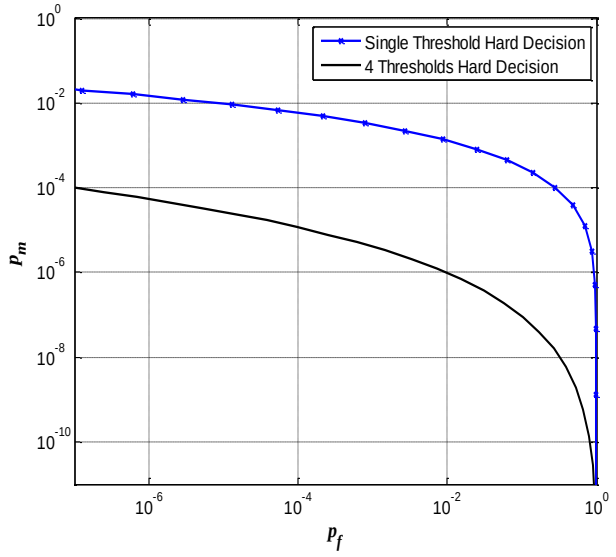


Figure 5. CROC for 4 thresholds hard decision combining technique compared to the single threshold, majority rule in the two cases,average $SNR = 20\ dB, m = 1.8, \Delta = 3, \Delta_c = 4, k = 5$.

 ISBN: 978-1-61208-203-5

VIII. CONCLUSION AND FUTURE WORKS

In conclusion, collaborative spectrum sensing model is studied in this paper when the sensing channel is a general Nakagami fading channel. Hard decision combining techniques were used to combine the collaborated sensors signal where each sensor shares only one or few bits to represent its sensing results. Single and multi-thresholds techniques are considered. Results show that both of them improved the system sensing performance significantly. It is also found that the multi-threshold works better than the single threshold especially in low SNRs. Results were compared to the other soft decision techniques used in the literature. It is found that soft techniques perform better but it has high fixed cost. On the other hand, the multi-level hard techniques are simpler to implement and have the flexibility tradeoff between the performance and the cost according to the application and the channel type.

REFERENCES

[1] http://www.portiodirect.com/productDetail.aspx?pid=49$55$51$431, Portio Research Mobile Factbook 2009, retrieved: April, 2012.

[2] http://www.sharedspectrum.com, retrieved: April, 2012.

[3] Fedral Communication Commission, "Spectrum Policy Task Force," Rep. ET, Docket No. 02-135, Nov. 2002.

[4] F. F. Digham, M.-S. Alouini, and M. K. Simon, "On The energy Detection of Unknown Signals over Fading Channels," IEEE Transactions on Communications, vol. 55, no.1, pp. 21-24, January 2007.

[5] A. Ghasemi and E. Sousa, "Opportunistic Spectrum Access in Fading Channels Through Collaborative Sensing," Journal of communications, vol. 2, no. 2, pp. 71-82, March 2007.

[6] H. Urkowitz, "Energy Detection of Unknown Deterministic Signals," Proc. IEEE, vol. 55, pp. 523-531, April 1967.

[7] H.Birkan Yilmaz, Tuna Tugcu, and Fatih Alagoz, "Uniform Quantizer for Cooperative Sensing in Cognitive Radio Networks", PIMRC 2010, September 2010.

[8] Q. Zou, S. Zheng, and A. H. Sayed, "Cooperative Spectrum Sensing Via Coherence Detection," 15th workshop on statistical signal processing, IEEE/SP, pp. 610 – 613, Aug. 31 2009-Sept. 3 2009.

[9] Y.-C. Liang, Y. Zeng, T. Hoang, and E. Peh, "Sensing-Throughput Trade-off for Cognetive Radio Networks," IEEE Trans. on wireless communications, vol. 7, no. 4, pp. 1326-1337, April 2008.

[10] A. Ghasemi and E. Sousa, "Fundmental Limits of Spectrum-Sharing in Fading Environments," IEEE Trans. on wireless communnication, vol. 6, no. 2, pp. 649-658, Feb. 2007.

[11] O. Al-Bashir and M. Abdel-Hafez, "Opportunistic Spectrum Access Using Collaborative Sensing in Nakagami-m Fading Channel with Real Fading Parameter," 7th International Wireless Communications and Mobile Computing Conference (IWCMC), pp. 472 - 476, July 5-8, 2011, Istanbul, Turkey.

[12] A. H. Nuttal,"Some Integral Involving the Q_M -Function," IEEE Trans. Inf. Theory , vol. 21, no. 1, pp. 95-96, Jan. 1975.

[13] A. Papoulis, "Probability, Random Variables, and Stochastic Processes," McGraw-Hill Europe; 4th edition (January 2002).

[14] C. Shannon, "A mathmatical theory of communication," Bell Systems Technical Journal 1948, 27, pp. 623-656.

[15] I. S. Gryadshteyn and IM. Ryzhik, "Table of Integrals, Series, and Products," 5th Edition. San Diego: Academic, 1994.

[16] W. Tranter, K. Shanmugan, T. Rappaport and K. Kosbar, "Principles of Communication Systems Simulation with Wireless Applications," Prentice Hall, 2004.

 ISBN: 978-1-61208-203-5

Analysis of Interfered Noise for Sound Systems over LTE Mobile Phones

Suna Choi, Sungwoong Choi, Sangbong Jeon, Yongsup Shim, Seungkeun Park
Spectrum Engineering Research Team
Electronics and Telecommunications Research Institute (ETRI)
Deajeon, Korea
sunachoi@etri.re.kr, swchoi@etri.re.kr, sbjeon@etri.re.kr, sys@etri.re.kr, seungkp@etri.re.kr

Abstract— **As the LTE is deployed commercially, the effect of interference of LTE mobile phone is issued. The paper presents an interference analysis of LTE mobile phone in the sound systems. Three LTE mobile phones from different manufacturer's brands are used for measuring interference and three GSM phones are tested for the relative comparison with the LTE phones. A speaker and a wire telephone are applied as sound systems which are affected by LTE phone. The output spectra exposed to the LTE phones are presented and the interfered noise levels in sound systems are experimented at various distances from the LTE phones, and also at various powers of the LTE phones. The experimental results show that LTE mobile phones generate an interfered noise in sound systems up to the distance of 30~40cm. Also, the interfered noise is generated at over the power of 0dBm.**

Keywords- LTE; interference; sound system.

I. INTRODUCTION

Evolution of wireless technology has been achieved in phase during a remarkably short time. The first generation (1G) has fulfilled the basic mobile voice, while the second generation (2G) has introduced capacity and coverage. This is followed by the third generation (3G), which has opened the gates for higher speed mobile broadband. The significant expansion seen in mobile and cellular technologies is a result of the increasing demand for high-data-rate transmissions [1], [2]. As the fourth generation (4G) cellular networks, which offers high performance and capacity, the long term evolution (LTE) has been proposed by 3rd generation partnership project (3GPP) [3].

Tens of countries already provide LTE mobile phone service and many others are preparing to start the service. Accordingly, the number of LTE user is expected to grow steeply. As the usage of LTE mobile phones is increased, the study on interference issue of LTE cellular phone with the sound systems is necessary.

LTE has adopted orthogonal frequency division multiple access (OFDMA) for downlink and single carrier frequency division multiple access (SC-FDMA) for uplink as the communication method [4]. In these methods, both of time and frequency division multiple access are employed to support multiple users. Therefore, the interference problem which causes the noise to sound systems can be emerged in LTE system as similar to a global system for mobile communication (GSM). Cellular telephone such as GSM is already known to cause electromagnetic interference with sound systems because of the pulsed nature of the signal of the time division multiple access (TDMA).

The interfered effect of GSM is investigated in many researches and applied to standards. In [5] and [6], the frequency spectra from several mobile phones including GSM and code division multiple access (CDMA) services are measured and compared. In [7], GSM modulation signal is suggested for the immunity test of sound and television broadcast receivers and associated equipment.

On the other hand, the interfered effect of LTE has not been investigated even though the need of interference analysis of LTE is growing. This paper focuses on interference analysis of LTE mobile phone in sound systems such as a speaker or a wire telephone, based on the measurement. LTE mobile phones from three different companies are used to investigate for the purpose. Additionally, three GSM mobile phones from identical companies are tested for the relative comparison with the results of LTE mobile phones.

The paper is organized as follows: Time domain structure of LTE system is described in Section II. Measurement methods are shown in Section III. Then, experimental results are presented in Section IV. Discussion is given consecutively in Section V. Finally, conclusions are followed in Section VI.

II. CHARACTERISTICS OF LTE SIGNAL

LTE supports two radio frame structures for frequency division duplex (FDD) and time division duplex (TDD) modes [8]. In the FDD mode, uplink and downlink transmission are separated in the frequency domain. In the TDD mode, uplink-downlink configurations with both 5ms and 10ms downlink-to-uplink switch-point periodicity are supported.

In this paper, we test LTE mobile phones which operate in the FDD mode, which has been deployed commercially. As shown in Fig. 1, the generic radio frame of LTE in FDD mode has time duration of 10ms in the time domain. A frame is divided into 20 slots of each 0.5ms. Each slot consists of a number of symbols. Although one slot composes the smallest resource block, the basic time-domain unit for scheduling in LTE is one sub-frame. Two consecutive slots form a sub-frame of 1ms duration. The

 ISBN: 978-1-61208-203-5

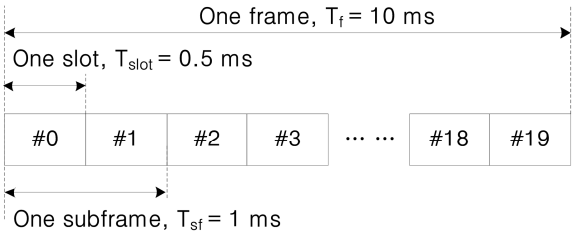


Figure 1. Frame structure of LTE

fundamental frequency associated with the LTE frame rate of 100 Hz and several of the harmonics fall in the audible frequency region. Therefore, they can cause interfered noises to sound systems.

III. MEASUREMENT METHOD

For the analysing the effect of interference of LTE, the interfered noise levels in sound system are experimented according to various distances from the LTE phones and various powers of the LTE phones. Three LTE mobile phones from different manufacturer's brands are used for measuring audible interference. Additionally, three GSM phones from identical brands tested for the relative comparison with LTE phones. In order to ensure the confidence, the experiment is progressed in the anechoic chamber.

As shown in Fig. 2, a base station emulator and an antenna are used to control the LTE and GSM phones. The powers of the LTE and GSM phones are adjusted from 0 to 22dBm as the commands of base station emulator. A speaker and a wire telephone are investigated as sound systems at various distances from the LTE and GSM phones. The interference-induced sound pressure levels (SPL) are recorded by a sound pressure meter.

The spectra of interferences from the LTE mobile phones can be visualized by an oscilloscope. The interferences are measured in the time domain and then transferred to the frequency domain signal using the FFT function of the oscilloscope. However, the measurement of sound pressure levels is performed without the oscilloscope to prevent the driving sound of the oscilloscope from affecting the sound pressure meter.

IV. EXPERIMETAL RESSULTS

A. Reference noise levels

As listed in Table 1, initial sound levels of the sound systems are measured and the results are used as the reference interference levels for the following measurements. The reference levels are performed when the LTE and GSM phone are off and the each power of a speaker and a wire telephone are on. Because the wire phone is measured with the receiver picked up, the measured reference sound level of the wire phone is about 5dB louder than that of the speaker.

Sound systems	Reference sound levels
Speaker	42.2 dB
Wire telephone	47.2 dB

Table 1. Reference sound levels of sound systems

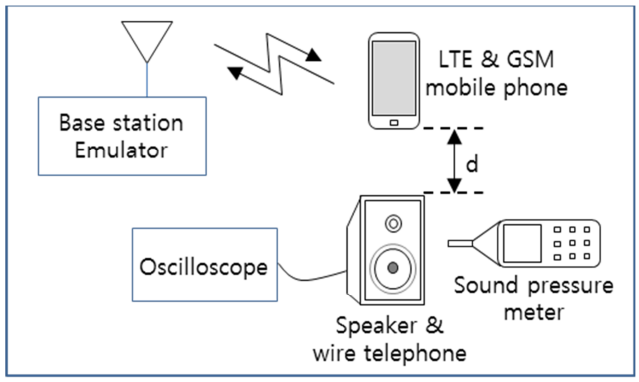


Figure 2. Measurement set-up

B. Spectra of interfered noise

The spectra of a speaker near to the LTE and GSM mobile phones are measured by wire connection from the output of speaker and the input of the oscilloscope.

Figs. 3 and 4 present the spectra of the interfered noise when the power of LTE and GSM mobile phones are 20dBm and the distance between each of LTE and GSM mobile phone and the speaker is 5cm. The interfered noise from LTE phone repeats at 10ms intervals and that from GSM phone repeats at 4.16ms intervals in the time domain. They match to the frame length of LTE and GSM. The peak voltage of the LTE phone is 1.19Vpp while that of GSM phone is 1.53Vpp.

Frequency spectra of Figs. 3 and 4 are presented from 0 to 20 kHz which is known as acoustic frequency range. The interference from LTE and GSM mobile phones reveals discrete peaks in the frequency domain. These peaks correspond to the frame rate and its harmonics. The type of interference produced by these technologies may be described as a buzzing sound in sound systems.

C. Interferences at the maximum power

The interfered noise levels of sound systems are examined when the power of the LTE phones is set to maximum value (22dBm) by the base station emulator to demonstrate the worst case of the interference scenario.

Plots of sound pressure levels of a speaker versus separated distance between the speaker and each LTE and GSM mobile phone at the maximum power are given in Fig. 5. The interfered effects of GSM mobile phones are tested for the purpose of relative comparison with LTE phones.

The initial distance between the speaker and each LTE and GSM mobile phone is set to 5cm, and then adjusted at intervals of 10 cm. Signs of A, B and C of Figs. 5 and 6 indicate the three brands of LTE and GSM phones. There are differences of measured sound levels among the LTE.

 ISBN: 978-1-61208-203-5

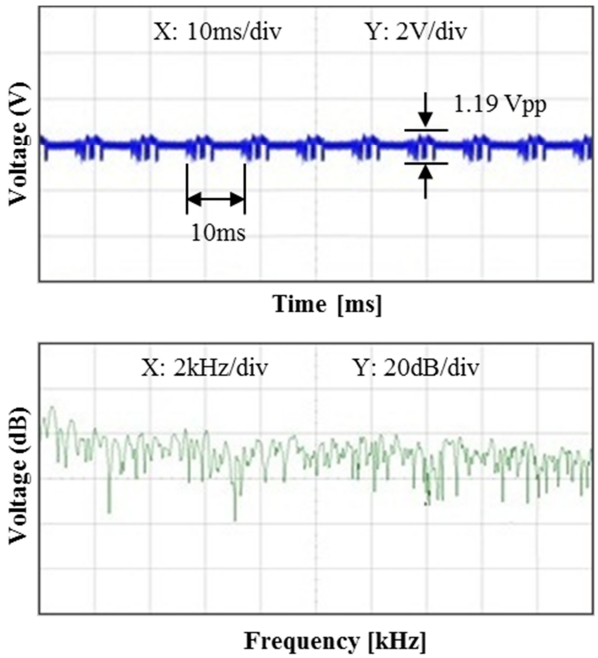


Figure 3. Spectra of interfered noise from LTE mobile phone

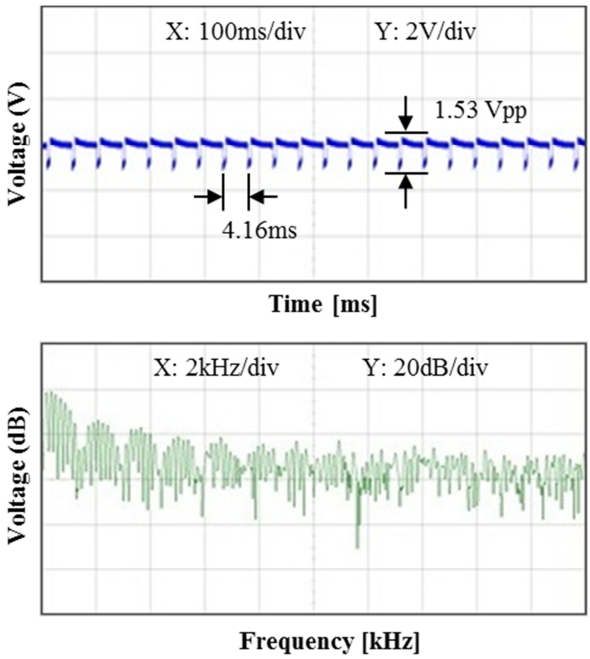


Figure 4. Spectra of interfered noise from GSM mobile phone

However, all of the interfered noises of the speaker from the LTE phones show the similar tendency. Even though the interfered noise levels from the LTE phones are lower than that from GSM phones, the interfered noises are detected up to a distance of about 25 cm. The interfered sound levels are approached to the reference value of Table 1 beyond 25cm.

The interferences of LTE mobile phones to a wire telephone at the maximum power (22dBm) are shown in Fig. 6. The initial distance between the speaker and each LTE and GSM mobile phone is set to 3cm, and also adjusted at intervals of 10 cm.

The interfered noise levels from the LTE phones are lower than that from GSM phones. While the interfered noises of GSM phones are detected up to a distance of over 63cm, the interferences of LTE phone diminishes and reaches to the reference value at a separation distance of almost 33 cm.

D. Interferences at various powers

The interfered noise levels of sound systems are investigated when the output powers of LTE mobile phones are varied from 22dBm to 0dBm for examining the effect of power.

Fig. 7 presents the plots of sound pressure levels of a speaker versus output powers of LTE mobile phones when the separation distance between a speaker and each LTE hone is fixed to 5cm. The interfered noise levels of the speaker decrease as the power of LTE phones declines. When the power reaches to 0dBm, the interfered sound levels are almost reaches to the reference values listed in Table 1.

Plots of sound pressure levels of a wire telephone versus output power of each LTE mobile phone are presented at Fig. 8 when the separation distance between wire telephone and LTE phones is fixed to 3cm. Although there are differences in the measured values among the LTE mobile phones, the interfered sound levels almost reach to the reference value at the power of 0dBm.

V. DISCUSSION

The audible frequency of human ear is generally known from the minimum 20Hz to the maximum 20 kHz. The experimental results show that LTE mobile phones apparently generate an interfered noise in the audible frequency region to sound systems such as a speaker and wire telephone, even though the interfered noise levels from the LTE phones are lower than that from GSM phones.

The interfered noise levels decrease as the distance between sound system and the LTE phone is increase. The interfered noise is almost vanished when the LTE mobile phone is about 30~40cm apart from the sound system. Therefore, a simple way to avoid the interfered sound noise is to place the LTE phone away from the sound system. But it cannot be a fundamental method for reducing the interfered noise.

The interfered noise levels also decrease as the power of LTE mobile phone declines. The output power of a LTE mobile phone in practical circumstance is determined by the distance from the base station and radio propagation environment. As the communication coverage of a base station is smaller, the output power of LTE mobile phone can be reduced. If the number of base stations is increased and the power of LTE mobile phone is reduced to less than

 ISBN: 978-1-61208-203-5

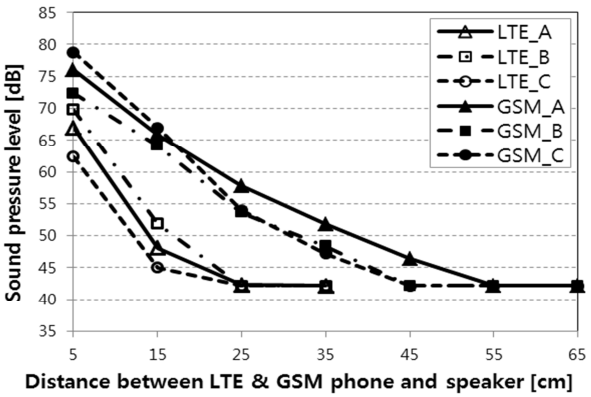


Figure 5. Interferences from LTE and GSM mobile phones for a speaker

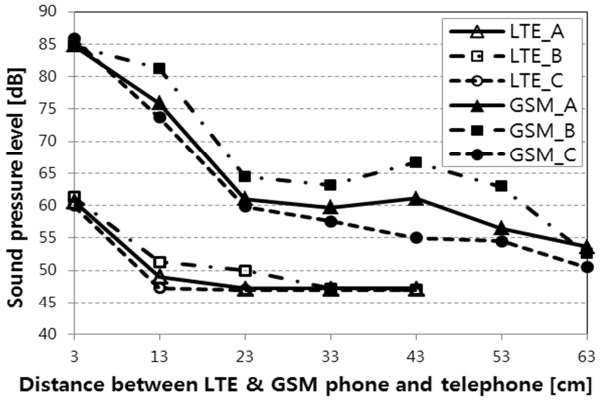


Figure 6. Interferences from LTE and GSM mobile phones for a wire telephone

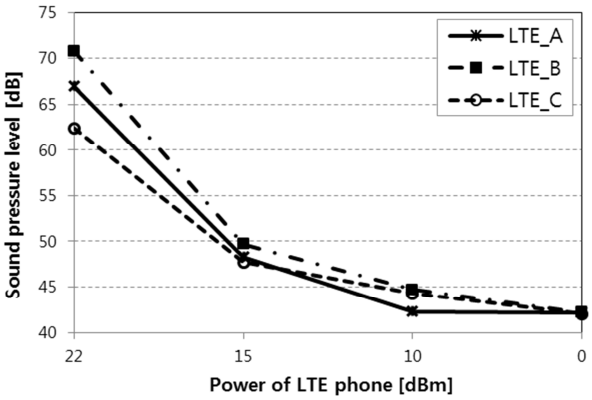


Figure 7. Interferences for speaker at various LTE powers

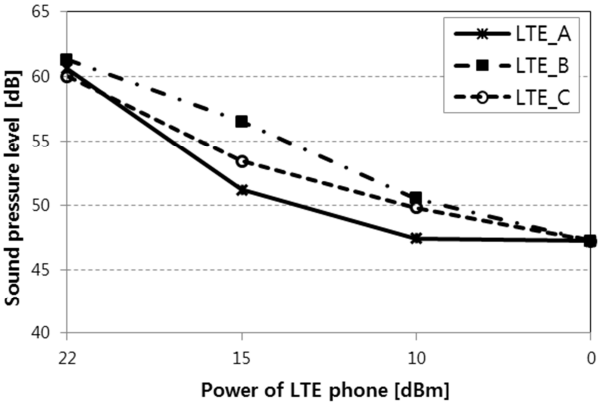


Figure 8. Interferences for wire telephones at various LTE powers

0dBm, the distance which the sound system is affected by a LTE mobile phone is expected to decrease to less than several centimetres.

VI. CONCLUSIONS

This paper presents an interference analysis of LTE mobile phone in the sound systems, based on the experiment. Three brands of LTE mobile phones investigated and a speaker and a wire telephone are applied as sound systems. The output spectra of interfered noise are presented in the time and frequency domain. Furthermore, the interfered noise levels in sound systems are measured at various distances from the LTE phones, and also at various powers of the LTE phones. The experimental results show that LTE mobile phones apparently generate an interfered noise to sound systems. Therefore, various efforts to reduce the interfered noises of LTE mobile phones such as separating from the sound systems or reducing the transmission power are necessary.

ACKNOWLEDGMENT

This research was supported by the KCC (Korea Communications Commission), Korea, under the R&D program supervised by the KCA (Korea Communications Agency) (KCA-2011-08921-01303)

REFERENCES

[1] M. Ergen, Mobile broadband - Including Wimax and LTE, Springer, NY, 2009

[2] E. Dahlman, S. Parkvall and J. Skold, 4G LTE/LTE-Advanced for Mobile Broadband, Academic Press: Elsevier, 2011

[3] D. Astely, E. Dahlman, A. Furuskar, Y. Jading, M. Lindstrom, and S. Parkvall, "LTE: the evolution of mobile broadband," *IEEE Commun. Mag*., vol. 47, pp. 44-51, Apr. 2008

[4] H.Holma and A.Toskala, LTE for UMTS-OFDMA and SC-FDMA based Radio Access, Wiley and Sons, 2009

[5] M. Skopec, "Hearing aid electromagnetic interference from Digital wireless telephones", *IEEE Trans. Rehab. Eng*., vol. 6, pp. 235-239, June 1998

[6] R. E. Schlegel and F. H. Grant, "Modeling the Electromagnetic Response of Hearing Aids to Digital Wireless Phones," *IEEE Trans. on EMC*, vol. 42, no. 4, pp. 347-357, Nov. 2000

[7] CISPR 20: Sound and Television Broadcast Receiver and Associated Equipment - Immunity Characteristics-Limits and Methods of Measurement, IEC, 6th Edition, 2006

[8] 3rd Generation Partnership Project, Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation, 3GPP Std. TS 36.211 V.10.3.0, 2011

 ISBN: 978-1-61208-203-5

Spatial Reuse and Interference-Aware Slot Scheduling in a Distributed TDMA MANET System

Isabelle Labbé, Jean-François Roy, Francis St-Onge, Benoit Gagnon
Communications Research Centre
Ottawa, ON, Canada
{isabelle.labbe, jean-francois.roy, francis.st-onge, benoit.gagnon}@crc.gc.ca

Abstract— **One of the goals pursued by this work is to gain a better understanding of the conditions for which spatial reuse in distributed TDMA ad-hoc networks is possible. Such understanding becomes particularly important when considering modern ad hoc networking. With the emergence of software programmable radios that support multiple modes of operations, the effects incurred by operating in low vs high spectral-efficiency mode should be well understood and ideally addressed by the protocol layers if system efficiency is to be preserved. A distributed TDMA system presented in [1] is re-visited to make use of an extended interference model. The extended interference model combines the graph-based interference model with the SINR-based interference model. A description of the cross-layering communication developed between the MAC and the PHY layers to support the model is given. The performance of the TDMA system is evaluated in simulation for both, the graph-based model and the extended model. The effect of the propagation environment (path loss exponent) and of the modulation requirement on spatial slot reuse is studied. Results show that network performance of the graph-based model rapidly degrades as the spectral-efficiency mode increases. The impact is even greater with decreasing values of the path loss exponent. In comparison, the extended model produces good performance results in all operating conditions.**

Keywords- spatial reuse, interference, slot scheduling, distributed TDMA Ad Hoc Network.

I. INTRODUCTION

Mobile Ad Hoc Networks (MANETs) have a continued growth in bandwidth demand mainly driven by the introduction of new user services and applications. A solution to providing increased capacity of wireless systems is to operate over wider bands so that more information can be sent. But because spectrum resources are limited and its usage restricted, this solution is not always possible and certainly not sustainable in the long term. An alternative approach in delivering increased capacity has been the development of high spectral-efficiency radios. High spectral-efficiency radios make use of advanced modulation techniques to transmit a higher capacity of bearer data without increasing the assigned channel bandwidth. The approach, however, is not without tradeoffs. The most important one being range. Operations at high spectral-efficiency modes will invariably reduce the achievable communication range. A strategy to compensate for the loss of range is to employ multi-hop network relaying. This approach of sacrificing range to the benefit of capacity (transmitting at high spectral-efficiency modes) while relying on relays to extend the coverage seems to be establishing in MANETs. This is the case in military tactical networks, for example, where there is an increasing need for more bandwidth to support the explosion of IP-centric operations and where multi-hop relay capability is very desirable to connect nodes that are temporarily out of range under terrain impediments or node movements.

In the past two decades, many protocols that address multi-hop capabilities in MANETs have been proposed. Amongst them, TDMA-based protocols have received much attention mainly because of their ability to provide QoS guarantees. An interesting characteristic of TDMA-based media access control (MAC) protocols is their potential for achieving higher network capacity through spatial reuse of the time slots [2]. Spatial reuse allows geographically separated nodes to schedule concurrent transmissions. The challenge of spatial reuse lies with the capability of generating an efficient scheduling algorithm that takes interference into account to prevent unnecessary message losses. Hence, an accurate modeling of interference is fundamental.

A large majority of the slot schedule designs (and thus of the slot reuse schemes) described in the literature have assumed a simple disk signal coverage model also known as the graph-based interference model [3-7]. In the recent years, the poor validity of the graph-based interference model and its unrealistic propagation representation has received much attention [8-14]. In all of those works, a more accurate physical interference model that uses the signal-to-interference-and-noise ratio (SINR) to describe the aggregate interference in the network is instead proposed. A comparison between the two interference models and their impact on network performance is presented in [8]. The simulation results show that in some cases, graph-based scheduling performance suffers when compared to interference-based scheduling. The study, however, does not consider various propagation models. The performance evaluations are presented for a specific path loss exponent value and for fixed communication and interference thresholds only. In [11-14], heuristic algorithms that build TDMA link schedules by taking into account the more accurate physical interference model are proposed. Most, if not all, lack presenting their work within the context of an actual protocol (i.e., as an integrated component). This leaves open important aspects of ad hoc networking such as

 ISBN: 978-1-61208-203-5

information distribution and conflict resolution. The problem is then formulated under simplified and/or unrealistic assumptions that undermine the practical relevance of the work.

In this paper, spatial reuse for distributed TDMA-based ad hoc networks is investigated. Several papers that consider both interference models present their work assuming a particular communication model in which the propagation parameters (e.g., radio power, SNR, path loss exponent) are set to the specific environment under study. Different from those, we take a generic approach to the characterization of spatial reuse. Our characterization tries to establish the conditions of operation for which a given interference model is valid. This is achieved by defining the set of parameters that have the greater impact on the interference models. Once identified, the conditions under which spatial reuse is deemed possible are derived for each model.

Based on the results obtained from the spatial reuse characterization, we present an extended interference model that combines the graph-based model with the SINR-based model. We validate the approach by integrating the proposed extended model into an actual prototype implementation of a TDMA-based MAC protocol [1]. An overview of the MAC-PHY cross layering approach used in support of the integration is provided along with the enhancements made to the distributed dynamic slot scheduling scheme. Using network simulation, we evaluate the performance of the TDMA system for various conditions of operation. In particular, we study the effects of operating the radios in low vs high spectral-efficiency modes. We also verify the impact of varying the path loss model. Performance results are presented for both the original protocol design (which was based on the graph interference model only) and the re-visited design (which is now based on the combined interference model).

II. NETWORK CONNECTIVITY MODELS

A. The Graph-based Interference Model

Most scheduling algorithms proposed for distributed TDMA-based multi-hop networks use a simplified binary propagation model. This model assumes a radio transmission range that stops at a finite border i.e., it assumes no or negligible residual energy beyond that border. Direct node-to-node connectivity (1-hop neighborhood) is possible for all nodes located inside a transmitterôs disk coverage.

In the graph model (such as shown in Figure 1), the interference from direct neighbors of a receiver is considered while cumulative interferences from nodes beyond 1-hop from the receiver are ignored.

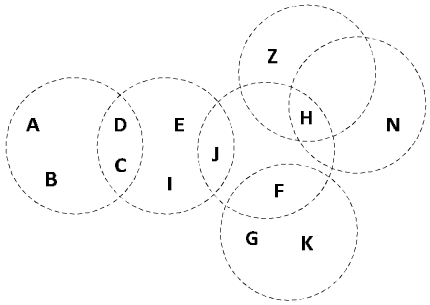


Figure 1: Simple disk-based network connectivity model

The MAC protocols elaborated under this model will typically try to maintain collision-free slot allocations by respecting the following conditions:

When traffic is intended for all neighbors (typically referred to as node scheduling), a communication from node I to all 1-hop neighbors is successful if no other node within node Iôs 2-hop neighborhood (in this case, nodes A, B, C, D,E, F, H and J) is transmitting in the same time slot as transmitter node I.

When traffic is intended for an individual neighbor (typically referred to as link scheduling), a communication from node J to receiver node I is successful if:
- Receiver node I and its 1-hop neighbors (in this case, nodes C, D and E) are not transmitting in the same time slot as transmitter node J;
- node Jôs neighbors (in this case, nodes C, D, E, F, H and I) are not receiving in the same time slot as transmitter node J.

Based on the above, a slot reuse schedule can be obtained for nodes that are geographically separated. For example, slot reuse for node-scheduled transmissions will be possible when transmitter nodes are separated by a distance of at least 3 hops. Similarly, slot reuse for link-scheduled transmissions will be possible between 1-hop transmitter nodes if their respective intended receivers are at least 3-hops apart. Such spatial slot reuse scheduling has been used by many distributed multi-hop TDMA MAC protocols to increase the capacity of the network and maximize the throughput [3, 5]. The drawback of this network connectivity representation is the over-simplification of the radio model by assuming that the signal of a transmitter node has no or negligible interference effect beyond a fixed propagation radius/range. This assumption may be valid under some specific conditions, as will be discussed further, but in many cases, this unrealistic representation of the propagation model may seriously impact the slot reuse scheme (and thus the overall capacity of the network).

B. The Physical Interference Model

An alternative and more accurate approach for achieving efficient spatial slot reuse is to consider the full interference environment i.e., to include in the connectivity model the

 ISBN: 978-1-61208-203-5

contributions of all received signals, namely the ones that are too weak to provide reliable communication but yet, can still cause a non-negligible interference. This model is known as the physical interference model [8, 9]. The physical interference model is based on signal propagation properties and the distance between the nodes. The SINR is used as a measure of the perceived network interference at a receiver node. A transmission is successfully received if the SINR at the receiver is higher than a given threshold.

To establish the conditions under which spatial reuse will be possible in the SINR interference model, we derive the minimal distance separation that must be respected between the main transmitter node and an interfering node (simultaneous tx) as a function of SINR values at the receiver. Figure 2 illustrates a possible node-scheduled slot reuse scenario valid under the graph-based interference model (since the transmitting nodes T and I are separated by a 3-hop distance).

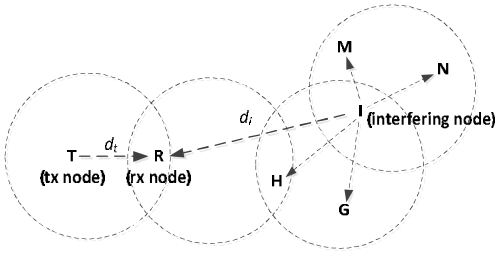


Figure 2: Interfering node scenario

Letôs assume that a signal transmission is going from a transmitting source node T to a receiver node R. The source node T is located at a distance d_t of the receiver node R. At the same time, an interfering node I located at a distance d_i from the receiver node R starts another transmission (intended for its own neighbors nodes G, H, M and N). Let P_t denote the power of the signal from the transmitter node T. In the absence of interference, it is generally accepted that the received power of a signal at the receiver is obtained as the ratio of the transmit power to the path loss. The path loss models the signal attenuation over the distance. Path loss is caused by the dissipation of power radiated by the transmitter as well as the effects of the channel propagation. The complexity of signal propagation makes it difficult to obtain a single model that characterizes path loss accurately across a range of different environments. We choose to use a simple model that captures the essence of signal propagation without resorting to complicated path loss models which are, in the end, only approximations of the real channel. Possible channel impediments such as multipath fading and shadowing effects are ignored. The formulation is derived based on the classical model for radio signal propagation in wireless networks. According to [15], the received power is modeled as:

$$P_r = P_t \left[\frac{\sqrt{G\lambda}}{4\pi d} \right]^{\alpha} \quad (1)$$

where P_r is the received power, P_t is the transmitted power, G is the gain of Tx and Rx antennas, ɔ is the wave length, d is the distance between the transmitter and the receiver and Ŭ is the path loss exponent. A path loss value Ŭ = 2 corresponds to the open space environment. The open space environment models an ideal environment for signal propagation. To account for attenuation due to ground or terrain effects, a path loss exponent value greater than 2 is generally used (typically 2 < Ŭ < 4). The higher the path loss exponent value, the greater the signal attenuation will be relative to the distance.

The SINR at receiver node R is defined as follows:

$$SINR = \frac{P_r}{P_i + N} \quad (2)$$

where P_r denotes the received power of the signal from the transmitter node T, P_i denotes the received power of the signal from the interfering node I and N represents the ambient noise at the receiver. Ignoring noise (since noise background is expected to be much lower than the interference signal) and combining (1) and (2), equation (3) is obtained:

$$SIR = \frac{P_t \left[\frac{\sqrt{G\lambda}}{4\pi d_t} \right]^{\alpha}}{P_i \left[\frac{\sqrt{G\lambda}}{4\pi d_i} \right]^{\alpha}} \quad (3)$$

We assume a homogenous ad hoc network where all nodes transmit at the same power (thus $P_t = P_i$) and at the same frequency. The successful reception of the signal sent by the transmitting node T depends on the *SIR* at node R. The signal is assumed to be valid (successful reception) if the *SIR* is above a certain threshold. After reduction, formula (3) becomes:

$$SIR = \left(\frac{d_i}{d_t} \right)^{\alpha} \quad \text{Ó } SIR_{threshold} \quad (4)$$

The relation of the interference range to the transmission range can thus be expressed as follows:

$$\frac{d_i}{d_t} \geq \left(\sqrt[\alpha]{SIR_{threshold}} \right) \quad (5)$$

Equation (5) was derived based on the linear path loss model. A more common way of expressing the measured

 ISBN: 978-1-61208-203-5

SIR is using a dB value. Equation(5) with the *SIR* value expressed in dB becomes:

$$\frac{d_i}{d_t} \geq \left(\sqrt[\alpha]{10^{SIR_{threshold}(dB)/10}} \right) \quad (6)$$

Equation (6) shows that for the reception to be successful, a minimum relative distance separation between simultaneous transmitting nodes must be met. This relative distance value depends on the desired SIR threshold and the particular path loss exponent of the propagation environment. The conditions for spatial reuse are thus determined by the relationship of the interference range (distance of the interfering source to the receiver node) to the transmission range (distance of the transmitting source to the receiver node). The minimum ratio requirement and thus spatial reuse conditions can be plotted for various values of $SIR_{threshold}$ and path loss exponents. Figure 3 shows the result for 4 $SIR_{threshold}$ values and 5 path loss exponent values.

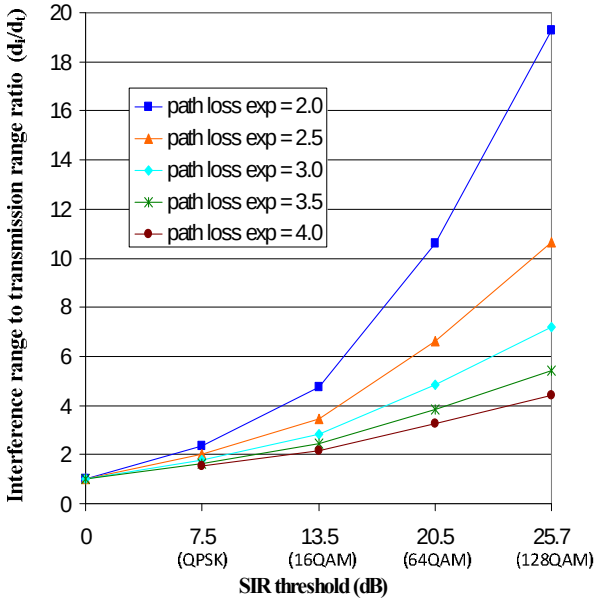


Figure 3: Min. relative distance separation vs SIR thresholds

To illustrate, the $SIR_{threshold}$ values selected for representation on the graph correspond to Signal-to-Noise (SNR) threshold values of an actual tactical VHF/UHF OFDM-based modem [1] operating at various modes over a 200 kHz bandwidth. For each mode of operation, the SNR threshold value corresponding to a Bit Error Rate (BER) of 10^{-6} was selected. A BER value of 10^{-6} is generally considered acceptable to obtain the full rate at the mode of operation. The SNR threshold values represented on the graph correspond respectively to coded modem rates of 195 kbps (QPSK), 390 kbps (16QAM), 653 kbps (64QAM) and 913 kbps (128QAM).

Figure 3 shows an interference to transmission range ratio which increases along with spectral efficiency. This relation implies that nodes transmitting with lower spectral efficiency are likely to achieve greater spatial slot reuse (since the minimum geographical relative distance requirement between simultaneous transmitting nodes is less). Consequently, the increase in network capacity gained from spatial slot reuse is expected to be higher when operating at a lower rate as opposed to higher rate modes.

The minimum relative distance requirement increases even more with decreasing path loss exponent values. For example, in the free space propagation environment (where path loss exponent value = 2), a relative distance node separation greater than 10 is required when operating at a 64QAM modulation mode. This ratio decreases to an approximate value of 4 when the path loss exponent rises to a value of 3.5. This impact of the path loss exponent is significantly reduced at lower SNR values. At QPSK for example, a path loss exponent variation of 2 to 3.5 causes only a small variation (1.5 to 2.3) of the corresponding distance ratio requirement. Lower spectral-efficiency modes are thus less affected by the propagation environment than higher spectral-efficiency modes.

It should be noted that the results presented in Figure 3 were obtained assuming only one source of interference. In a typical ad hoc network, contributions are likely to come from multiple sources of interference. In such cases, the resulting aggregation of all signals at the receiver will impact the distance required for spatial reuse which will inevitably increase. Results derived from eq (6) thus constitute a best case scenario.

C. *Limitations of the Graph-based Interference Model*

We now consider the minimum relative distance requirement in the context of the graph-based interference model. As previously stated, the graph-based interference model ignores the physical reality of RF propagation. The model imposes a static spatial separation between simultaneous transmitter nodes which does not always meet the minimal distance ratio requirement necessary to produce collision-free spatial reuse schedules. To better understand the issue, a simple case scenario is illustrated in Figure 4.

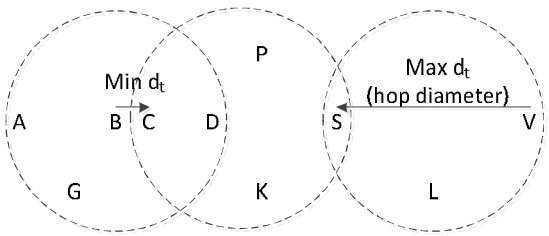


Figure 4: Relative distance separation in graph-based model

 ISBN: 978-1-61208-203-5

For node scheduled transmissions, the graph-based interference model imposes a spatial separation of at least 3-hops between simultaneous transmitter nodes. In Figure 4, this means that nodes A, B or G can simultaneously share slots with nodes L or V without causing any collision at C, D or S. In reality, this will be true only if the minimum relative d_i/d_t ratio is respected for the required SNR threshold value. Letôs consider the case where the d_i/d_t ratio is maximal. The d_i/d_t ratio will be at its maximum when the interference source is located as far as possible from the receiver node while the transmitter node is located as close as possible to it. In Figure 4, this takes place for example, when node B is transmitting to node C and interfering node V is transmitting to node S. The resulting ratio at node C is $(2*hop_{diameter})/\min d_t$. If the distance of the transmitter is small compared to the hop diameter, the resulting ratio value will be large enough to ensure that no collision occurs at node C (regardless of the SNR threshold value). To validate this spatial slot reuse scenario, the resulting d_i/d_t ratio must equally be measured at receiver node S. Node V is now the transmitter node while node B becomes the interfering node. The resulting ratio at node S is $(\min d_t + hop_{diameter})/ hop_{diameter}$. Keeping the assumption that $\min d_t << hop_{diameter}$, this results in a ratio of ~ 1. According to Figure 3, this low ratio value will inevitably produce a collision at receiver node S, regardless of the SNR threshold value. Thus, maximizing the ratio on one side has the effect of minimizing it on the other. This behavior seriously reduces the efficiency of the slot reuse scheme.

This simple case scenario illustrates well the limitations of the graph-based interference model. To meet the relative distance criteria, the model tends to require some sort of symmetry in the relative nodes location. This goes against the very nature of ad hoc networking. Obviously, some cases exist where the criteria will be satisfied. However, in most of those cases, the resulting d_i/d_t ratio at the receiver nodes will likely be relatively low. Based on those observations, it is reasonable to expect sub-optimal performance results from a slot reuse scheme that would strictly be based on the graph-based interference model.

Since the ratios derived in Figure 3 are relative separation distances as opposed to absolute distances, brief considerations should be made regarding the physical limitations imposed by the curvature of the earth. It is well known that the line-of-sight (LOS) communication range between two points is limited by the horizon and depends on the height of the antennas at each point. From [16], the LOS distance in kilometers can be computed as:

$$dist_{LOS} = 8.24\left(\sqrt{h}\right) \quad (7)$$

where h is the height of the antennas (assuming identical transmit and receive antennas) in meters. From (7), the maximum expected LOS distance between two nodes for antenna heights of 3m and 20m is of 14km and 37km respectively. The former corresponds to a fair estimate of the maximum LOS distances between ground-to-ground mounted vehicles while the latter is representative of ship-to-ship communications at sea. Beyond this distance the nodes cannot see each other and thus the radios cannot interfere with one another. The effect of the earthôs curvature must therefore be taken into consideration when deriving the minimum distance required for allowing slot reuse. Clearly, it can put an upper bound on the results presented in Figure 3 and in some cases, preserve the validity of the graph-based interference model.

The analysis presented in this section has shown the limitations of representing network connectivity based on the simplified disk signal coverage model. The analysis has revealed that the model can be used in support of spatial slot reuse but only under some specific conditions of operation. In particular, the model is expected to provide some throughput increase when the radios are operated in low spectral-efficient modes (because of the lower distance ratio requirement). It is also expected to perform well when the transmission range is large (in which case, the physical limitation due to the earth curvature comes into play and preserves the validity of the interference model). When operating outside of these conditions, the model starts to suffer significantly from distant node interference (border effects), affecting network performance and the ability to perform efficient slot reuse.

III. THE EXTENDED INTERFERENCE MODEL

Previous work carried out by the authors in the area of distributed TDMA ad hoc networking, has led to the design and development of an experimental prototype system called the *MATRIQS* [1]. *MATRIQS* is a distributed TDMA-based multi-hop system developed to provide enhanced tactical IP networking capabilities within battle group units. Designed to be flexible and adaptive, the *MATRIQS* system supports programmable VHF/UHF waveforms with multiple modes of operation. Various degrees of spectral-efficiency modes are offered with data rates ranging from 9600 bps (low efficiency, low bandwidth, high robustness mode) to 1.0 Mbps (high efficiency, high bandwidth, low robustness mode). Currently supported bandwidths are 25, 50, 100, 200 and 350 kHz.

The *MATRIQS* MAC developed initially and presented in [1] automatically achieved spatial slot reuse based on the traditional graph interference model only. When characterizing the system in various operating conditions, the limitations of the interference model and its impact on the system performance were observed. A more realistic network connectivity model was needed. The approach adopted to address the problem was to extend the graph model to include physical interference considerations. Essentially, the approach that we propose is a combination of the two interference models. The concept is to keep the simplified disk coverage model to establish the first level of

 ISBN: 978-1-61208-203-5

interference knowledge. Then, the more accurate physical (SINR) interference model is applied to the slots that are identified as potentially available for reuse by the protocol (as an outcome of the first level). The slot scheduling/slot reuse scheme resulting from this combined two-step approach has the benefit of remaining efficient and accurate through a wide range of operating conditions while keeping the implementation complexity at an acceptable level.

Conceptually, the approach is similar to the hybrid solution presented in [17]. However the two methodologies differ greatly. The algorithm proposed in [17] uses an iterative scheme based on a fixed interference range value. The interference range is increased adaptively and new squared conflicted graphs are re-generated until an interference-free schedule is found. No distribution aspects are discussed and the algorithm implies global network connectivity knowledge. Our scheme, instead, relies on the distribution of slot information to dynamically guide the slot allocation decisions. While the reported slot information makes use of the graph-based interference model to ensure distance-2 non-conflicting node scheduling, it also includes physical interference information that ensures interference-free slot reuse scheduling. This solution not only maintains the increased capacity provided by spatial slot reuse but it also preserves the flexibility of the protocol in terms of dynamic slot allocations.

An overview of the modifications that were performed to the *MATRIQS* MAC protocol to support the extended interference model is provided next.

A. The Cross-Layering Approach

The physical interference model makes use of the SINR to evaluate the perceived network interference at a receiver node. Since this specific channel information can only be obtained by the physical layer (modem), a cross-layering communication approach was developed between the MATRIQS MAC layer and its underlying modem.

The cross-layering exchange between the two layers occurs via abstract generic interfaces that conform to Software Defined Radio (SDR) principles. The communication enables the MATRIQS protocol to derive a ñper slotò channel quality value which is used in the protocolôs slot scheduling and allocation algorithm.

The ñper slotò channel quality is expressed as a binary value. The value is either 0 or 1, where 0 indicates a ñgoodò slot with low rx interference level and 1 indicates a ñbadò slot suffering from high rx interference level. To derive this channel quality, the MAC obtains, at the end of each slot, two parameters from the modem: the rx signal power (S) and the noise + interference power (N), as measured and estimated by the modem for the slot period. The rx signal power can only be measured by the radio frequency receiver. Receivers contain an automatic gain control (AGC) device used to normalize the output signal level. The control voltage (V_G) of the amplifier is derived from the input signal level and follows a known transfer function. This signal can be supplied in digital form and be used to derive the absolute incoming signal power. The RF input signal power (S + N) is calculated as follows:

$$P_{RF} = P_D / f(V_G) \quad (8)$$

where P_{RF} is the incoming RF power (S + N), P_D is the power of the digitized signal after AGC (measured by demodulator) and $f(V_G)$ is the AGC transfer function and represents absolute gain. Since the MAC requires S and N to be separate values, the burden falls onto the demodulator to measure the noise (N) and therefore provide both S and N separately. In the event where the demodulator is unable to detect an incoming signal, it declares N = (S+N) where S = 0.

Using the rx signal power (S) value obtained from the modem, the MAC protocol maintains a run-time table of received power for each of the nodeôs 1-hop neighbors. The rx power value is averaged over a time window to smooth out the effect of possible transient conditions. The MAC then combines this information with its knowledge of slot status and ownership to compute an SINR value for each slot. The SINR value is calculated as follows: if the slot status is *rx*, the MAC first determines the slot ownership (i.e., which neighbor the slot belongs to). The MAC then extracts from the table the latest recorded rx signal power for that neighbor node and derives the SINR by using the ratio formula (S/N). If the slot is available (i.e., the slot does not belong to anyone), then no corresponding rx signal power value will be found in the table. The calculation of the SINR value cannot be performed at this point since it requires a relative comparison of a neighborôs rx signal level to the measured interference. In this particular situation, the worst-case approach is adopted. The MAC identifies from the table the node for which it has the weakest signal (lowest recorded rx signal power). The MAC then uses this value to compute the SINR for the slot.

For each slot, the channel quality is obtained by comparing the computed SINR value with the SINR threshold (typically set to correspond to a BER of 10^{-6}) for the modulation and error correction code in use. This channel quality estimate is provided by the physical layerôs demodulator. Here, the implication is that a matching good/bad signal threshold value must be pre-established and included in the programming. The channel quality for the slot is declared good if the computed SINR value is greater than the SINR threshold. It is declared bad otherwise.

The cross-layering communication enables the MATRIQS protocol to obtain and maintain a run-time per slot channel quality value that takes into considerations the full interference environment. The MATRIQS slot scheduling and allocation scheme was modified to take advantage of this channel quality information. The enhancements done to the scheme are described next.

 ISBN: 978-1-61208-203-5

B. The Distributed Dynamic Slot Allocation Scheme

The *MATRIQS* protocol supports a fully dynamic slot scheduling and allocation scheme [1]. The scheme is based on slot request and release. As for most distributed-based schemes, it combines two approaches: a pro-active approach and a re-active approach.

The pro-active approach makes use of the information readily available to guide decisions on selecting/requesting the slots that have the highest probability of producing error-free transmissions. The idea is to pick non-conflicting transmission allocations in the first place. Nodes request slots based on the distributed slot information they maintain. Each node reports slot ownership information at minimum once per cycle. The reported slot status information ensures that nodes request non-conflicting node-scheduled transmission allocations over a 2-hop neighborhood while taking advantage of simultaneous link-scheduled transmissions whenever possible. Spatial slot reuse based on the graph interference model is thus inherently supported by the protocol and may take place when transmitters (in the case of node-scheduling) or receivers (in the case of link-scheduling) are separated by a distance of at least 3 hops. As neighborhood slot information is collected, a node derives and maintains a set of slots it considers available for request. Essentially, this set includes all the slots that have been reported with the *available* status by the neighbors. A node selects the slots to request from that set.

The protocol also supports a re-active approach. The re-active approach offers a mean to bring corrections when problems or conflicts are detected. Conflicts may rise from sudden changes in conditions due, for example, to node mobility. The protocol implements the re-active approach by specifying a comprehensive conflict detection and resolution scheme. Actions/decisions resulting from this scheme typically translate into nodes issuing slot preemptions or objections to slot requests.

Because these design approaches were originally based on the graph interference model, the slots considered available by the protocol were often unusable due to the interference coming from remote nodes. Consequently, in many situations, decisions taken by the protocol to perform slot reuse led to an increase in the number of collisions and yielded sub-optimal performance. The channel quality information obtained from the cross-layering was included in both approaches (pro-active and re-active) to improve the slot allocation and scheduling scheme.

B.1 Extended Pro-active Approach

The pro-active approach is extended by including the channel quality value in a nodeôs periodic slot status report for slots that are advertised as *available*. To keep the overhead low, each slot status is expressed using a 3-bit code. Originally, only 2 of the 3 bits were used to indicate slots status *available*. The third bit is now used to report the slot channel quality value where a bit value of 0 indicates a ñgoodò slot (slot is considered interference-free) while a bit value of 1 indicates a ñbadò slot (a strong enough interfering signal has been detected in the slot).

This supplementary information is now used to refine the *available for request* slots set maintained by a node. The set now only includes the slots reported as *available good* by each neighbor. As a result, when making a request, a node will pro-actively select slots that are truly interference-free at that time. It is important to note that no penalty is paid in additional overhead cost. The cost lies with the increased complexity in the structure of the cross-layering solution.

B.2 Extended Re-active Approach

Because operating conditions of ad hoc networks vary over time (e.g., topology changes, propagation characteristics changes), slot schedules that were collision/interference free may suddenly not be anymore. The reactive approach is extended by considering the channel quality in the conflict resolution scheme. The parameter is integrated to guide slot preemption decisions. Originally, the slot preemption mechanism was strictly used as a means to resolve slot scheduling (ownership) conflicts. The mechanism is now also used to notify a sending node of bad slot receptions. A node will now also issue a preemption message to a neighbor for which signal reception on specific slot(s) has fallen below the SNR threshold. On reception of a preemption message, those ñbadò slots are immediately released by the transmitting node.

In the same manner, the channel quality value is now also considered within the slot approval process. Nodes now verify their latest slot channel quality values before approving or objecting to a request. This is because the interference conditions may change between the time the original slot selection is done and the time a neighbor makes its approval or objection decision. An approval is sent if no slot ownership conflict is found and the slot channel quality is good. An objection is sent otherwise.

IV. PERFORMANCE EVALUATION

To evaluate the performance of the system and to measure the impact of the enhanced scheme on slot reuse, the MATRIQS experimental system was ported into the QualNet (QN) simulation framework [18]. Most of the MATRIQS protocol stack (i.e., the MAC and link layers) was preserved during the porting process. Consequently, discrepancies between the actual system implementation and the simulated version are minimal. The QN physical abstract layer was used in place of the actual modem. It was modified to support the SINR-based interference model as well as the cross layering communication scheme.

 ISBN: 978-1-61208-203-5

A. Simulation Setup

In section 2, it was determined that spatial reuse is mainly affected by the propagation path loss and the modulation requirement (SINR threshold). Hence, a multi-hop network topology was constructed to which we applied various combinations of path loss exponent values and modulation modes (SINR thresholds).

To be consistent with the analysis presented in section 2, the QN radio signal propagation model was set to the classical log-distance path loss model. Scenarios were run for three path loss exponent values of 2.5, 3.0 and 3.5 respectively. Those values were selected to be representative of various types of propagation environments with a degree of attenuation ranging from mild to severe. For each path loss exponent value, the impact of the modulation was evaluated by varying the SNR threshold values. The SNR thresholds were 7.5 dB, 13.5 dB and 20.5 dB each corresponding to modulation modes QPSK, 16QAM and 64QAM respectively. The channel bandwidth was 200 kHz. The operating frequency was set to 300 MHz which is representative of low band UHF tactical operations. The resulting raw channel rates were 195 kbps (QPSK), 390 kbps (16QAM) and 653 kbps (64QAM).

The network topology was composed of 20 nodes deployed using the random uniform distribution. In order to evaluate the effect of both the propagation path loss and the modulation mode on spatial reuse, it was required that the d_i/d_t ratio values remained the same in all scenarios. This meant keeping the node layout and the relative distance between the nodes the same for all scenarios. This was achieved by scaling the size of the grid and by adjusting the transmission power level accordingly. For each scenario, once fixed, the transmission power was kept uniform and configured the same for all nodes. The resulting topology had a network connectivity diameter of up to 8 hops.

The network was fully connected at all time (there was always a path between any pair of nodes). To ensure worst-case interference, the network was saturated i.e., each node always had traffic to send. The traffic was UDP/CBR and sent by the MAC protocol using the node-scheduled transmission mode.

B. Results and Discussions

For each scenario, the following performance metrics were collected:

- reception collision ratio (%): total number of rx collisions over the total number of rx signals (locked signals)

- successful slot usage ratio (%): total number of successful slot tx (i.e., no neighbor rx collision) over the total number of slots

- network throughput (kbps): total number of bits successfully received in the entire network over the time period.

The results are presented in Figures 5, 6 and 7 respectively. The results refer to metrics averaged over 10 runs, each initiated with a different simulation seed. The simulated time was long enough to ensure that steady-state conditions had been reached. For each simulation run, data collection began only after the network was ñupò to avoid transient effects due, for example, to initial empty neighbor tables. For comparison purposes, the performance results are presented for both, the original protocol design (which was based on the graph interference model only) and the re-visited design (which is now based on the combined interference model).

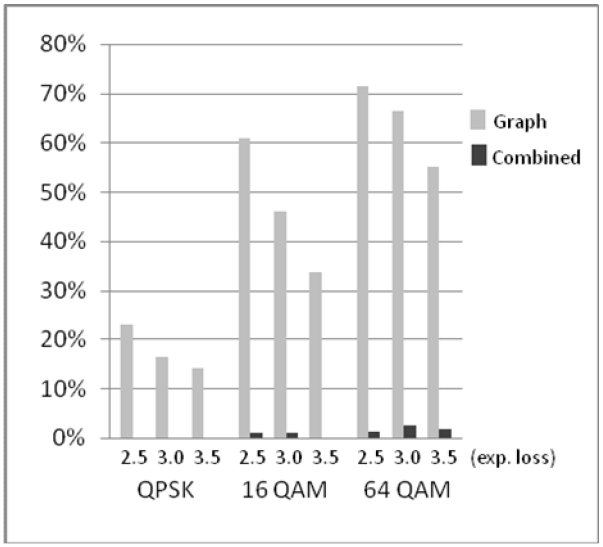


Figure 5: Rx collision ratio

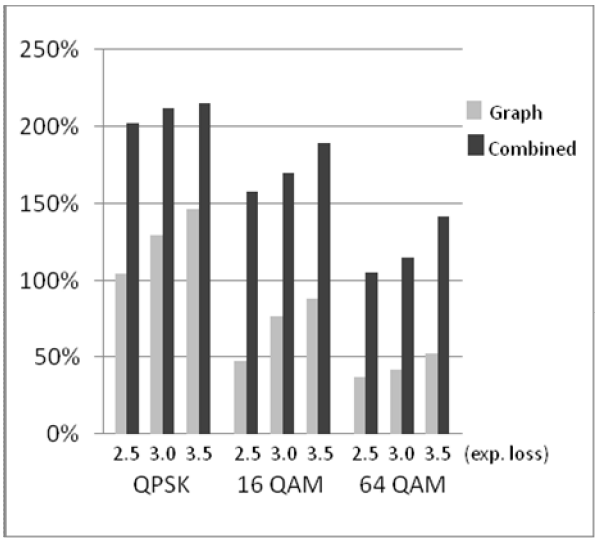


Figure 6: Successful slot usage ratio

 ISBN: 978-1-61208-203-5

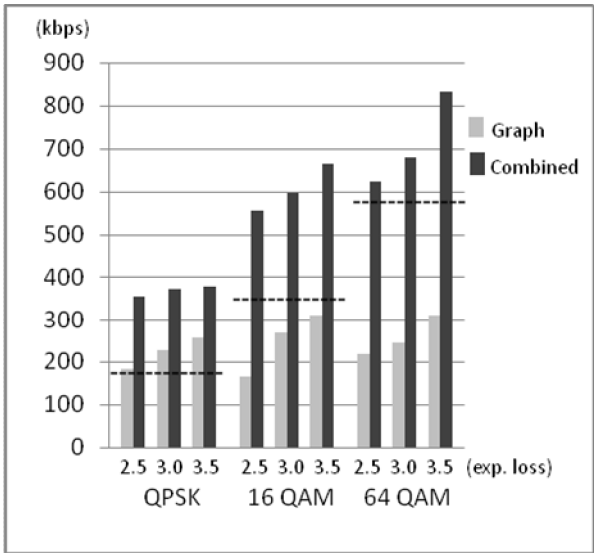


Figure 7: Network throughput

As expected, the performance results obtained with the combined interference model are better than the results obtained with the graph interference model only. A significant difference is seen with the measured reception collision ratios (Figure 5). While the protocol based on the combined interference model achieves quasi collision-free schedules, the graph-based protocol exhibits rx collision ratios between ~15% to ~70%. Figure 6 shows that the successful slot usage ratios obtained with the combined model surpass by as much as 70% to 110% the ratios measured with the graph-based model. For example, in the case of 64QAM+exp. loss 3.5, the graph-based model achieves successful transmissions in only half (50%) of the slots. In contrast, the combined model displays a successful transmission rate of 140%. In addition to achieving successful transmissions in all (100%) of the slots, an additional 40% is gained through the occurrence of simultaneous transmissions due to slot reuse (slot reuse ratios are represented by the portions in excess of 100% in Figure 6). Slot reuse translates into a direct increase in network capacity. This can be observed in Figure 7. Whenever slot reuse is present in Figure 6, Figure 7 shows corresponding network throughput values above the 100% user data capacity limit (indicated by the dotted line for each modulation mode). Thus any value in excess of the displayed thresholds represents a gain in network throughput resulting from the slot reuse scheme. Due to the success of the slot reuse scheme, the network throughput values (displayed in Figure 7) are consistently higher for the combined interference model. In some cases, a network capacity up to 3 times that of the graph-based model is obtained.

The performance results obtained for both approaches closely match the predicted behaviors of section 2. As expected, a higher ratio of spatial slot reuse and thus a greater increase in network capacity is achieved when operating at low spectral- efficiency modes (e.g., QPSK) than when operating at high spectral-efficiency modes (e.g., 64QAM). This is because the minimum distance ratio requirement is less for lower rate modes. It is thus more easily satisfied in the MANET (ref. Figure 3: min d_i/d_t = ~2 for QPSK vs min d_i/d_t values between 3 and 11 for 64QAM). It should be noted that for the graph model, slot reuse occurs in the QPSK modulation mode only. It takes place however, in all scenarios for the combined model.

While the results obtained with the combined interference model remain quite acceptable in all cases, the performance of the graph-based approach quickly degrades as the spectral- efficiency mode becomes higher. The impact is even greater with decreasing values of the path loss exponent. Clearly, as the minimum distance ratio requirement becomes greater, the graph-based model fails at meeting the criteria and its performance seriously starts to suffer. The amount of rx collisions increases drastically and the network throughput collapses as the successful slot usage ratios fall below 100%.

V. CONCLUSION

In this work, spatial reuse for distributed TDMA-based ad hoc networks is investigated. The simulation results confirm that the accuracy of the interference model is essential to fully benefit from a maximum network capacity. The simple graph-based interference model, for example, shows sub-optimal but yet acceptable performances when tested with low spectral- efficiency modes (e.g., QPSK). However, the validity of the model rapidly degrades as the spectral-efficiency mode increases. The reception collision ratio increases drastically and seriously impacts the network throughput.

The extended interference model proposed in this work clearly outperforms the simple graph-based approach. More importantly, it produces good performance results in all operating conditions. The approach requires the support of cross-layering communication between the MAC and the PHY layers but the resulting improvements are sufficient to justify the increase in complexity.

One of the goals pursued by this effort was to gain a better understanding of the conditions for which spatial reuse in distributed TDMA ad-hoc networks is possible. Such understanding becomes particularly important when considering modern ad hoc networking. With the emergence of software programmable radios that support multiple modes of operation, the effects incurred by operating in low vs high spectral-efficiency mode need to be well understood and ideally addressed by the protocol layers if system efficiency is to be maximized.

ACKNOWLEDGEMENTS

This work was supported by Defence R&D Canada (DRDC).

 ISBN: 978-1-61208-203-5

REFERENCES

[1] Labbe, I. et al., ñAn Adaptive VHF/UHF System for the Next Generation Tactical MANETs**",** in Proc. of MILCOM 2009, Oct. 2009.

[2] Nelson, R. and Kleinrock, L., ñSpatial TDMA: A collision-free multihop channel access protocolò, IEEE Trans. Commun., vol. COM-33, no. 9, pp. 934ï 944, Sept. 1985.

[3] Jorgenson, M. et al., ñOperation of the Dynamic TDMA Subnet Relay System with HF bearersò, in Proc. of MILCOM 2005, Vol 1, Oct., 2005.

[4] Bao, L. and Garcia-Luna-Aceves J. J., "A New Approach to Channel Access Scheduling for Ad Hoc Networks", in Proc. of ACM MobiCom 2001, July 2001.

[5] Young, C.D., "USAP Multiple Access: Dynamic Resource Allocation for Mobile Multihop Multichannel Wireless Networking", in Proc. IEEE MILCOM 1999, Oct. 1999.

[6] Kanzaki, A. et al., ñDynamic TDMA Slot Assignment in Ad Hoc Networksò, in Proc. of AINA 2003, March, 2003.

[7] Fan Y., Biswas, S.**, "**A Self Reorganizing MAC Protocol for Inter-vehicle Data Transfer Applications in Vehicular Ad Hoc Networks,**"** in Proc. of ICIT 2007, Dec., 2007.

[8] Gronvist, J., Hansson, A., ñComparison Between Graph-Based and Interference-Based STDMA Schedulingò, in Proc. of MobiHoc pp. 255-258, 2001, Oct., 2001.

[9] Gupta, P., and Kumar, P. R, ñThe capacity of wireless networksò, in IEEE Transactions on Information Theory 46, 2, March 2000.

[10] Xingang, G., Roy, S., Conner, W. S., ñSpatial Reuse in Wireless Ad-hoc Networksò, in Proc. IEEE VTC 2003, Fall 2003.

[11] Brar, G., Blough, D. and Santi, P., ñComputationally Efficient Scheduling with the Physical Interference Model for Throughput Improvement in Wireless Mesh Networksò, in Proc. of MobiCom 2006, Sept., 2006.

[12] Fan, S., Zhang, L., ñLink Scheduling with Physical Interference Model for Throughput Improvement in Wireless Multi-hop Networksò, in Proc. of CSIE 2009. pp.430-434, 2009.

[13] Gronkvist, J., ñDistributed Scheduling for Mobile Ad Hoc Networks ï a Novel Approachò, in Proc. International Symp. On Personal, Indoor and Mobile Radio Communications, pp.964-968, 2004.

[14] Brar, G., Blough , D., and Santi , P., ñThe SCREAM Approach for Efficient Distributed Scheduling with Physical Interference in Wireless Mesh Networksò, in Proc. Distributed Computing Systems, 2008.

[15] Goldsmith, A., ñWireless Communicationsò, Cambridge University Press, ISBN-10: 0521837162, 2005.

[16] Ryan, M.J., and Frater, M.R., ñTactical Communications for the Digitized Battlefieldò, Artech House, ISBN 1-58053-323-x, 2002.

[17] Christophides, F. and Friderikos, V. ñIterative hybrid graph and interference aware scheduling algorithm for STDMA Networksò, Electronics Letters, Vol 44, Issue 8, April 2008.

[18] http://www.scalable-networks.com/products/qualnet/

 ISBN: 978-1-61208-203-5

Vehicular Networks Smart Connectivity

Sivakumar Sivaramakrishnan and Adnan Al-Anbuky
Sensor Network and Smart environment Research Centre,
Auckland University of Technology,
Auckland, New Zealand
ssivakum@aut.ac.nz, aalanbuk@aut.ac.nz

***Abstract*—Good connectivity helps in reducing the number of packet drop and improve network efficiency. Vehicular networks have a pattern of flow and a defined heading. Connectivity break occurs due to insufficient radio range. Dynamic coverage modulation varies the range to maintain coverage. In areas of high vehicle traffic, data packet loss would increase due to packet collision. To avoid this the coverage of node could be reduced and data handoff would be activated for maintaining connectivity. Shorter coverage in dense areas allows multiple nodes to communicate with minimal interference. Unnecessary transmissions lead to network congestion, adaptive sampling collects route information to determine its suitability for establishing communication. This work attempts at minimizing network congestion and reduce breaks in connectivity through dynamically modulated coverage, direction dependent data handoff and adaptive sampling. The algorithm has been simulated using OPNET and results reflect reduced packet drop.**

Keywords-Adaptive connectivity; Data Handoff; Dynamic coverage modulation; Adaptive sampling.

I. INTRODUCTION

Vehicular networks or VANETs are networks in which vehicles communicate with each other to disseminate vital information regarding the vehicle and traffic. This includes vehicle malfunctions, closed lanes, speed limits, accidents, caution measures due to varying driving conditions and others. These networks aim at making roads safer by warning the drivers of any potential hazards on the road.

Data dissemination in the network depend on vehicles heading, and the density of vehicles in a given region. The heading of the vehicle as shown in Figure 1 is used to determine the suitability of the vehicle to receive and/or carry forward data. For example, if lane is closed for north bound traffic due to road work, the traffic traveling on the opposite lane (south bound) would carry the information. This in turn will pass it on to the north flowing traffic. The other capability available in vehicles these days is a predefined route to the destination accessed through GPS (global positioning system). Using this information, the vehicles heading can be predicted and information can be shared.

Communication in vehicular networks as discussed in [1], [2] takes place in the following manner.

- Vehicle to Vehicle (V2V)
- Vehicle to Infrastructure (V2I)

Information broadcast and repetitive data transmission lead to network congestion and increase collision. Leontiadis and Mascolo [3] propose time to live to expunge information from the network after certain duration of time in a specific area. The issue with this is determination of the time as it is scenario specific. It is dependent on traffic movement, if the traffic is moving slow the information would also need to live for a longer duration. On the other hand faster traffic requires shorter life cycle for the data. We propose route dependent and priority based time to live. This approach for determining the life cycle of data classifies different road hazards by the level of severity and assigning priority [4]. Zhao and Cao [5] use multihop to transmit data for sparse networks. The issue with the approach is that if data is not completely transferred, while the vehicle changes its heading, and deviates from the predicted path, the data would get lost.

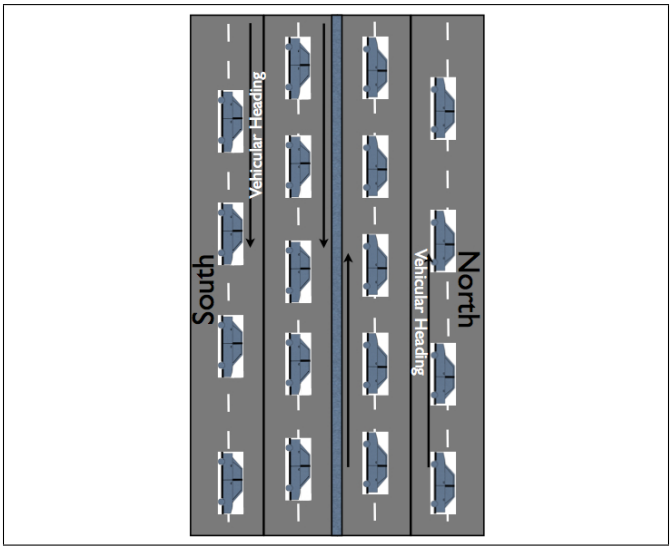


Figure 1. Definite Headings of Vehicles

Limiting the life cycle of high priority data in dense slow moving traffic does not serve the purpose. Hence, we propose modulating the radio coverage of the transmitter so that radio pollution is contained reducing communication

 ISBN: 978-1-61208-203-5

collisions and packet retransmissions. Modulation of radio coverage would allow for reducing the coverage in dense traffic and increase coverage in places of light traffic. Radio signals are also associated with a given maximum range. Moreover, the dynamics of the roads keep changing. It is therefore required to have data handoff to transfer large data packets over multiple nodes (vehicles). Multiple hop and data handoff both require multiple nodes but in data handoff communication set-up is only done once with the first receiving vehicle. The first receiving vehicle provides the details of the transmitting vehicle to the next receiving vehicle. This approach does not require setting up the communication with other receiving vehicles. This scheme has importance in scenarios where traffic on a lane is moving fast and on the other lane the traffic is slow as shown in Figure 2a.

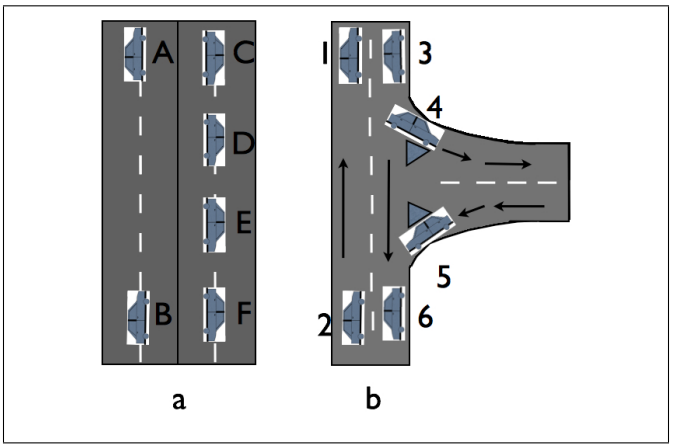


Figure 2. Traffic Scenario

The vehicles moving fast have a shorter time span to complete a communication. With large data packets the possibility of incomplete data transfers is high. To overcome this issue, if vehicle A (Figure 2a) handsoff data to vehicles F,E,D and C (in this order), the communication window of time gets wider, allowing for data packet to be successfully delivered. Figure 2b shows a scenario where direction of motion of the vehicle changes. As vehicle 4 in Figure 2b changes the path, data transmitted by vehicle 1 and 2 is not of any relevance, where as vehicle 5 is joining the lane and hence would need the information. Incorporating dependence of direction on handoff would reduce the re-transmission, thereby reducing packet collisions.

The paper is organized in the following sections. Section II discusses the effect of different parameters on vehicular connectivity. Section III gives the design for improving connectivity through handoff. Results are discussed in Section IV. Section V concludes the paper.

II. Factors Affecting Vehicular Connectivity

Connectivity amongst vehicles gets affected by multiple factors like vehicle velocity, transmission range, vehicle heading and route to destination.

A. Vehicle velocity

Velocity of a vehicle has a direct impact on the number of packets transmitted. At high velocities, the transmission time window reduces, which reduces the number of packets delivered. Assuming that the communication is specified by a packet size of x bytes and a bandwidth of B bps, then the required transmission time would be $T_{trans} = x/B$. For a vehicle moving at velocity v km/hr or $v/3.6$ m/s, the time required to cover a distance $d/1000$m is given as $T_{req} = d/v$. As communication takes place between two or more vehicles, hence their relative velocities are considered. For vehicle having the same heading (moving in parallel to each other) $v = v_1 - v_2$ and vehicle moving in opposite direction as shown in Figure 3, its $v = v_1 + v_2$. The communication

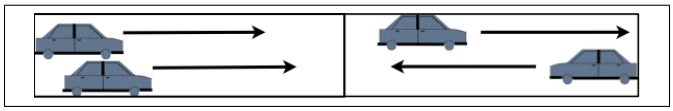


Figure 3. Effect of Velocity dependent on heading

is successful if $T_{trans} = T_{req}$, as this would allow complete packets to be delivered. Figure 4 shows that as the relative velocity of the vehicles increase, the amount of data packets received drops.

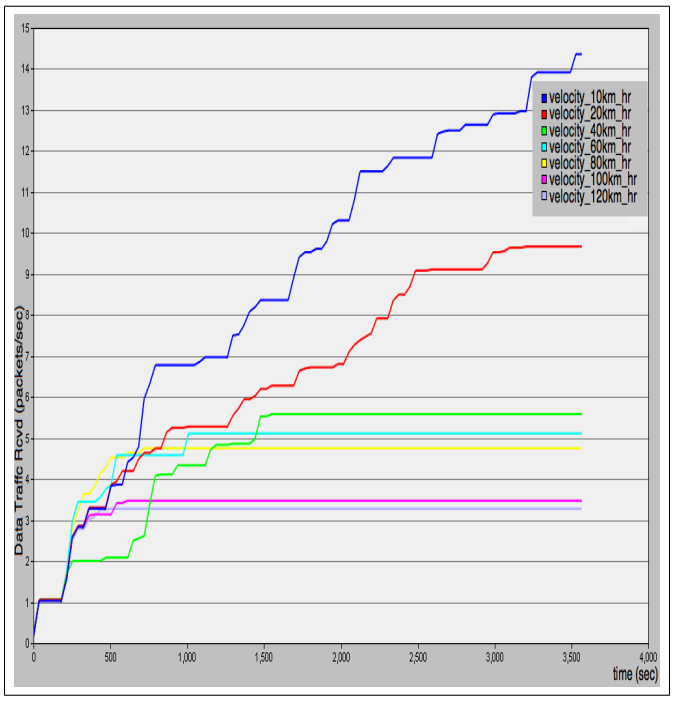


Figure 4. Available window of time for data communication against vehicle velocity

B. Transmission range

Invariably, vehicle velocity keeps changing with changing driving conditions. There are areas where vehicles move

 ISBN: 978-1-61208-203-5

slowly and traffic builds up creating a dense network of communicating objects. Somewhere else, the traffic may be free flowing and the network becomes sparse in communicating objects. Modulating the communication range through transmission power relative to the speed will help in keeping the window of required time constant.

C. Vehicle heading and route to destination

The heading of a vehicle plays an important role in determining the suitability of a vehicle to participate in communication. If in a heavy traffic zone all vehicle start broadcasting information there would be enormous data packet collisions resulting in failed communication. Most vehicles are equipped with GPS these days. These devices pre-calculate the route to the destination. Sharing this information with neighboring vehicles would help them decide if they have any vital information that needs to be shared. Even though GPS is so prevalent, but many drivers do not always rely on them due to their personal preferences [6]. Heading of the vehicle which depends on streets and lanes allows other vehicle to determine its suitability for data handoff. The uncertainty due to human involvement can be mitigated by coupling GPS information with the current heading for handoff. If a vehicle with which handoff has been performed, changes its heading suddenly, then it would necessitate retransmission, which is undesirable. The

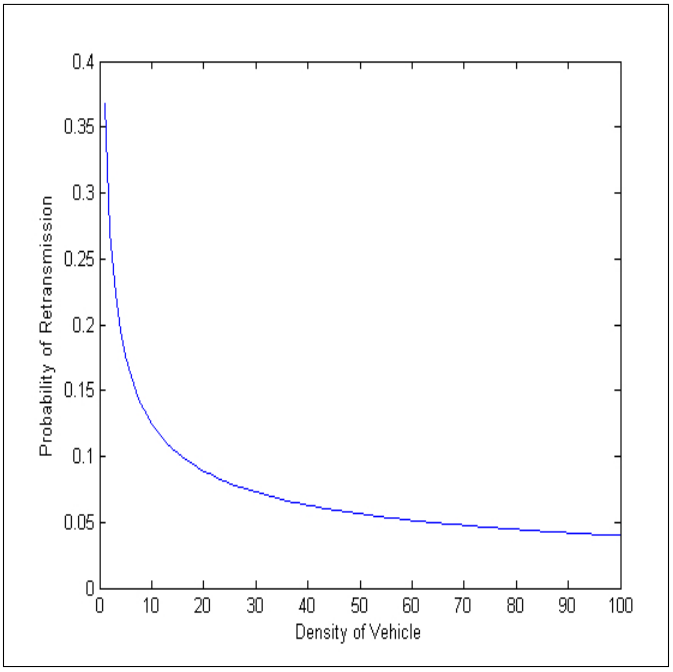


Figure 5. Probability of retransmission

dynamics on the road is stochastic and Poisson's distribution can be used to find the probability of retransmission.
$P = \frac{e^{-\lambda}\lambda^k}{k!}\ (k = 0, 1, 2, 3...)$
$\lambda =$ traffic density $\rho * a$(area)

$$P = \frac{e^{-\rho a}\rho a^k}{k!} \quad (1)$$

It can be inferred from the probability analysis, as shown in Figure 5, that the probability of retransmission is higher for areas where vehicle density is sparse as compared to high density areas. This is because the redundancy for handoff is higher in dense areas and therefore requirement of retransmission reduces.

D. Time to Live for Data Packets

Time to Live or TTL of a data packet is used to determine the length of time for which a data packet should remain in the network (with data packet being transmitted from one node (vehicle) to another). The issue is determining how long a data packet should be present before removing it from the network as setting a random value might remove the packet either too early or congest the network with redundant data. This work proposes to adapt the time to live based on the information priority $I_{priority}$, vehicle velocity v and route $GPS_{coordinates}$ as $TTL(I_{priority}, v, GPS_{coordinates})$.

III. Design for Improving Connectivity through Direction Based Handoff

The different factors in Section II contribute to maintaining connectivity. Based on these, the transmission coverage, requirements of handoff and TTL of data depend on independent factors like velocity, data size, vehicle heading and route information. The flow diagram in Figure 6 shows the trigger conditions for Coverage, Handoff, Direction and Adaptive Sampling (finding new vehicles to establish communication).

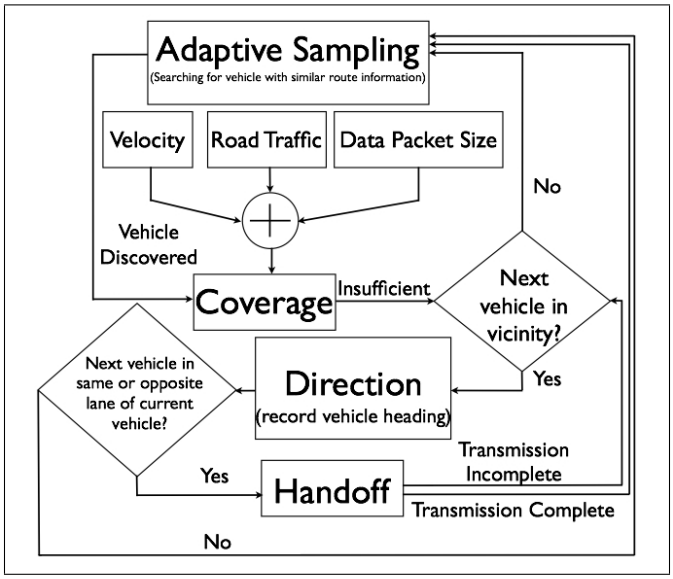


Figure 6. Flow Diagram for Connectivity of Vehicular Networks

1) Adaptive Sampling: Adaptive Sampling uses sensing coverage sense vehicles in the vicinity and exchange route information. It helps the vehicle wishing to transmit data to determine whether the other vehicle will move on the route for which the information is pertaining. If the vehicle is not following the route, then establishing the communication will be wasted. As sensing coverage also involves communication which results in use of radio resources, it should not be performed very often as it contributes to network congestion. When a vehicle has information pertaining to a particular area then it needs to know the possibility of the current route being used to reach that destination. Adaptive sampling helps in achieving this through the use of artificial neural network. It uses the past sensing information to predict the use of the current route to reach the area of interest. Depending on the outcome of the neural network the vehicle performs sampling using adaptive coverage. This reduces the sensing operation, minimizing use of radio and therefore reducing network congestion. Figure 7 shows the block diagram for adaptive sampling of the network to search for vehicles for data transmission.

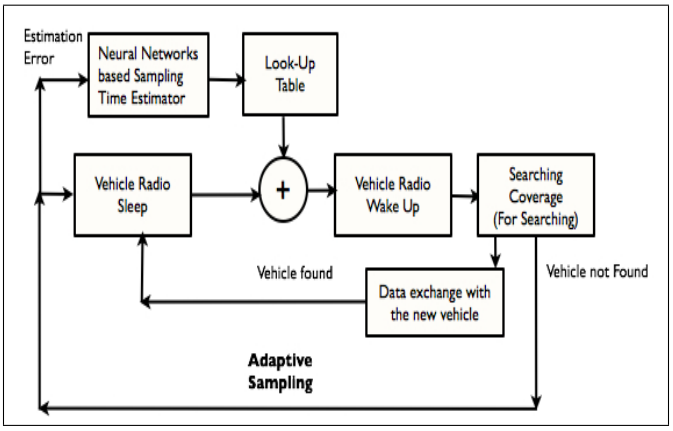


Figure 7. Adaptive Sampling for Vehicle selection

Figure 8 shows the packet format for the sensing operation. This reflects the key factors like destination, heading, route (start and end coordinates) and Hazard code (code, priority and traffic condition). The route coordinates give the start and end coordinates of the current route till the next intersection. It is not sufficient by itself because a driver may suddenly take a different route of their preference to the destination. It is therefore necessary to know the current heading and destination. As there can be multiple paths leading to the same destination, the vehicles heading will allow the determination of future coordinates. This is based on the use of current coordinates and heading to determine the route to destination. Route coordinates provide current information which would be important for the other vehicles. From Figure 7, it can be noted that artificial neural network estimates T_n (duration after which next sampling should be performed). In applications like VANETS assuming a constant traffic flow on the road would be unrealistic. For such circumstances, a single prediction of T_n would not be accurate. Hence, the output of the neural network is stored in a look-up table with possible values. When an estimation error is reported, a different value is taken from the table. The network re-trains once the values in the table are exhausted.

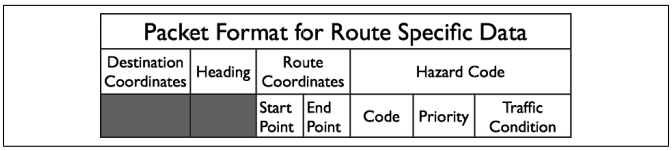


Figure 8. Packet Format for Vehicle selection

2) Adaptive Coverage: Constant radius of coverage does not provide the flexibility to selectively disseminate data to other vehicles. Adaptive coverage varies according to the density of vehicles on the road. For sparse density of vehicles, the coverage increases and reduces where density of vehicles is high. The advantage of this scheme is that it reduces packet drop and network congestion.

$$R_{\max} = \frac{(v \cdot T_{trans})}{2} \quad (2)$$

Here, R_{max} is the maximum radius of coverage at relative velocity v. On equating Equation 2 with the range equation (Equation 3)

$$R_{\max} = \sqrt{\frac{P_t G_t G_r \lambda^2}{(4\pi)^2 \, \mathrm{S}_{\min}}} \quad (3)$$

transmission power P_t is computed and the range is dynamically varied. Here G_t, G_r are transmitter and receiver gain respectively, λ is the wavelength and S_{min} is the receiver sensitivity. The algorithm for adaptive coverage is as follows.

Table I gives the algorithm used to determine nodes coverage.

Table I: Algorithm for Sufficiency of Coverage

Algorithm for Sufficiency of Coverage
: Sense the density of vehicle in the vicinity : Compute data size to be transmitted : Compute duration of connectivity T_{trans} : Sense Vehicle velocity

3) Direction: The direction or heading of the vehicle helps predict the future route of the vehicle. Figure 9 a,b,c and d show different routes between the same source and destination. Upto point A, all the four routes are same; after point A, the route remains same for Figure 9 a and d, but for b and c the heading changes.

Similarly, at point B, the route is same for Figure 9 a

 ISBN: 978-1-61208-203-5

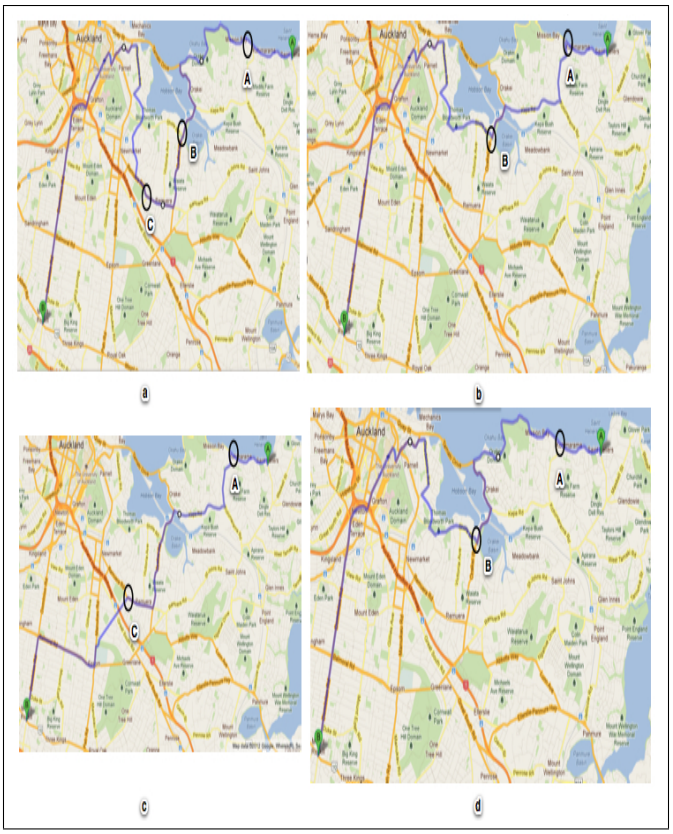


Figure 9. Vehicle heading and final destination

and c, whereas the heading changes for b and d. Similarly at point C heading changes between Figure 9 a and c. As destination is know the GPS device can calculate the different paths to the destination at each change in heading. If the vehicle which has some hazard information for a particular stretch of source, destination pair and the vehicle senses another vehicle heading along the same direction information handoff can take place.

4) Handoff: Handoff is required when the coverage is not sufficient because of vehicle velocity and limited coverage in high density regions. In situations like these, multiple vehicle having same heading (as relative velocity would be less for vehicles having same heading) are used to collaboratively collect the data and share amongst themselves to reconstruct it as shown in Figure 2a.

The handoff requires the vehicle transmitting V_{trans} data to transmit its Id, velocity, direction of motion, data packet size. Using this information, the vehicle receiving V_{rec} data inform their next hop vehicle about the transmitting vehicle V_{trans}, as shown in Figure 10.

As the vehicles are prepared to receive the data from the transmitting vehicle hence we save on packet exchange to establish the communication. Figure 11 shows the flow of the handoff sequence amongst receiving vehicle. The vehicle

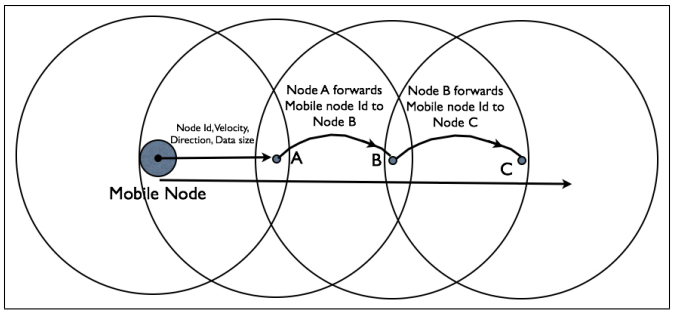


Figure 10. Handoff Scenario

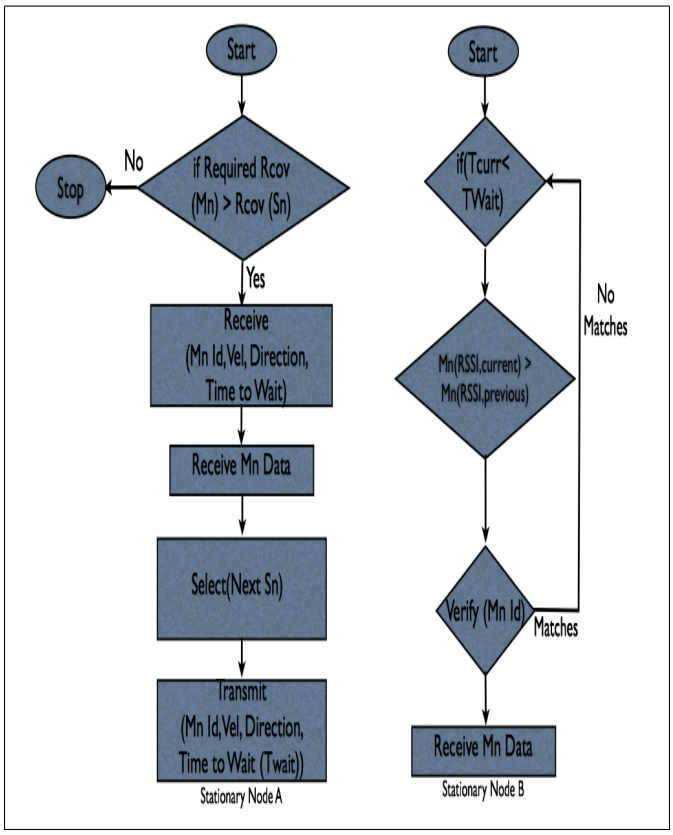


Figure 11. Handoff flow diagram

which is first to communicate with the transmitting vehicle M_n forwards its details to the next vehicle, which becomes ready to receive data from the transmitting vehicle. The Handoff flow checked by the receiving vehicle S_n checks, whether the coverage area of the transmitting vehicle is larger than its own and could data be completely transmitted. If it cannot be completely transferred, then the first receiving vehicle passes on the identification (M_nId, its velocity Vel, Direction and Time to Wait for the next node to receive data) of the transmitting vehicle to the next vehicle.

IV. Results and Discussion

Network congestion happens when data traffic is not controlled. Reduction in redundant data reduces the con-

gestion. Adaptive sampling associates route with data and determines the need to perform a transmission. Figure 12 shows the performance with and without Adaptive sampling. It is evident that the number of packets in the network has dropped leading to reduction in network congestion.

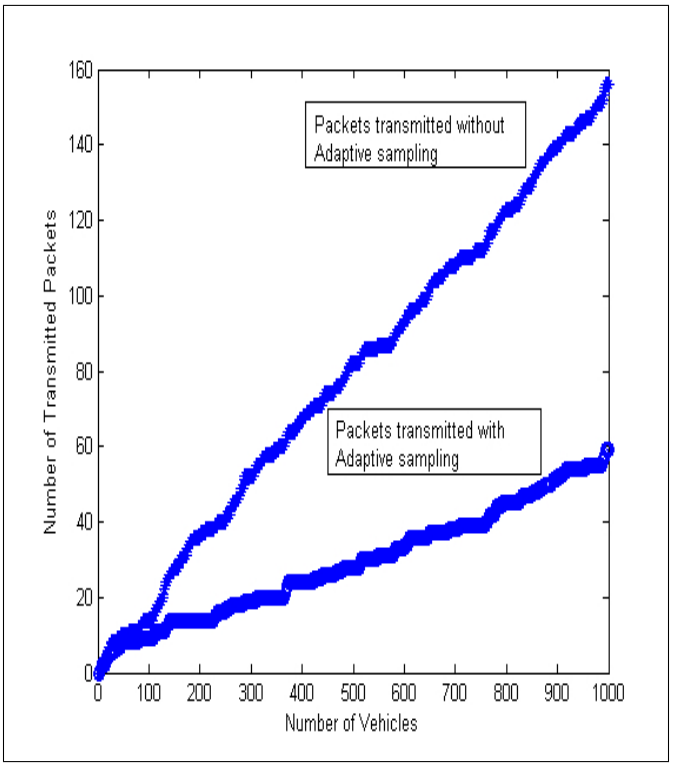


Figure 12. Number of Packets transmitted : Adaptive vs Regular sampling

Adaptive Coverage varies its coverage area depending on the traffic and the data size. Figure 13 shows that by adapting the coverage there is a reduction in the number of packets lost. The light blue and the green graphs are the number of packets transmitted and received for fixed coverage. It can be seen there is a significant packet loss. The red and the dark blue graph are for adaptive coverage. There is a significant reduction in loss of packets. In areas where the density of vehicles is less or in high density region where there is reduced coverage and a large packet needs to be transmitted, data handoff is performed to prevent break in connectivity. It's noted from OPNET simulation, Figure 14 that through handoff complete data is being transmitted over multiple vehicles, overcoming the issue of insufficiency of coverage.

Incorporating direction of motion to select a vehicle for data handoff, further reduces the unwanted data transmission. Implementing the algorithm within the nodes in OPNET shows that the vehicle which is on a lane changing its heading does not participate in communication. The result shown in Figure 15 graph 3 shows that without implementation of direction data is being received as the vehicle is

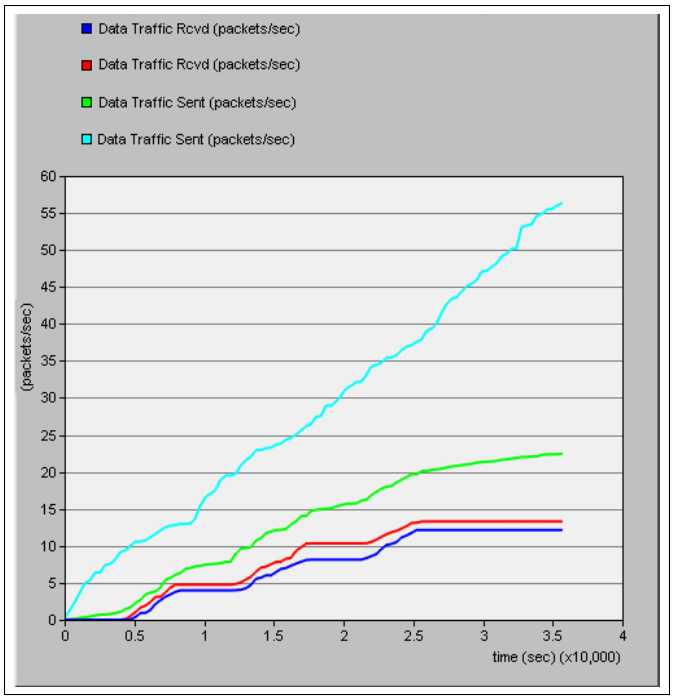


Figure 13. Comparison of Packets transmitted and packets received for Adaptive Coverage versus Fixed Coverage

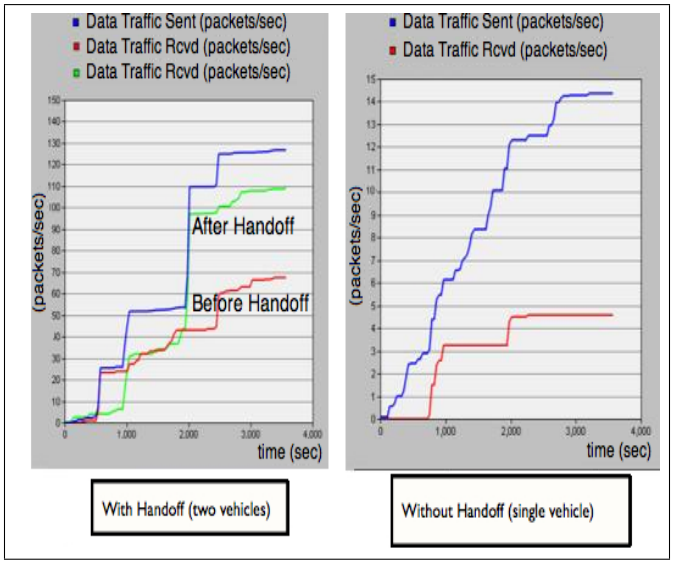


Figure 14. Comparison of Handoff with Coverage

in communication range but with the implementation of direction even though the vehicle is still in communication range but data is not received. The other vehicle graph 2 moves in a lane parallel to the transmitting vehicle graph 1 and hence receives data.

 ISBN: 978-1-61208-203-5

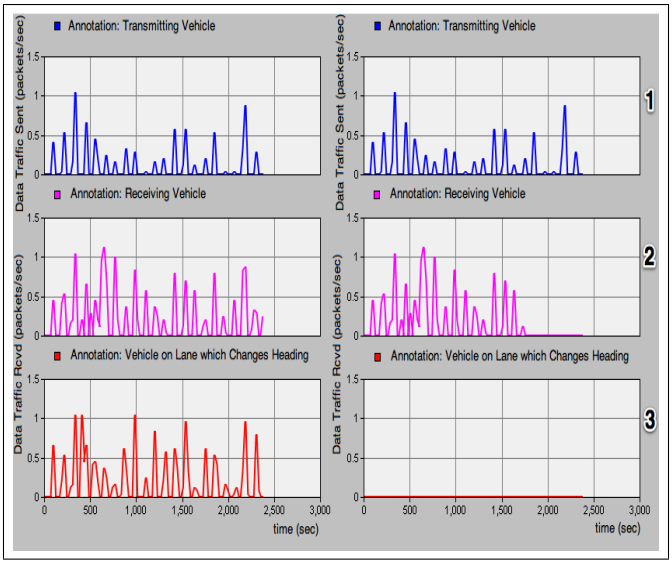


Figure 15. Direction Independent vs Direction Dependent transmissions

V. Conclusion

The work presented here utilizes Adaptive sampling, Adaptive coverage, Direction of motion of Vehicle and Data Handoff for intelligently sensing the need for a packet transmission. This approach reduces the number of redundant packet transmission and hence network congestion. The results show that through adaptive sampling a significant decrease in unwanted transmissions is achieved. Adaptive coverage modulates the transmission power to harness the opportunity and maintain the connectivity. It is evident from the results that in comparison to the fixed radius coverage, there is very small packet loss in adaptive coverage. Data handoff is used to maintain connectivity in scenarios where coverage is not sufficient for transmitting large data packets. Improved throughput is evident in the results. Implementation of direction allows handoff to perform selective dissemination of data. This paper shows that network congestion and packet loss is reduced through controlling the coverage radius, data dissemination and vehicular collaboration through handoff.

References

[1] Y.-T. Chang, J.-W. Ding, C.-H. Ke, and I.-Y. Chen, "A survey of handoff schemes for vehicular ad-hoc networks," in *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference*, ser. IWCMC '10. New York, NY, USA: ACM, 2010, pp. 1228–1231. [Online]. Available: http://doi.acm.org/10.1145/1815396.1815677

[2] H. Menouar, M. Lenardi, and F. Filali, "Movement prediction-based routing (mopr) concept for position-based routing in vehicular networks," in *Vehicular Technology Conference, 2007. VTC-2007 Fall. 2007 IEEE 66th*, 30 2007-Oct. 3 2007, pp. 2101 –2105.

[3] I. Leontiadis and C. Mascolo, "Opportunistic spatio-temporal dissemination system for vehicular networks," in *Proceedings of the 1st international MobiSys workshop on Mobile opportunistic networking*, ser. MobiOpp '07. New York, NY, USA: ACM, 2007, pp. 39–46. [Online]. Available: http://doi.acm.org/10.1145/1247694.1247702

[4] M. Bouassida and M. Shawky, "On the congestion control within vanet," in *Wireless Days, 2008. WD '08. 1st IFIP*, Nov. 2008, pp. 1 –5.

[5] J. Zhao and G. Cao, "Vadd: Vehicle-assisted data delivery in vehicular ad hoc networks," in *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, April 2006, pp. 1 –12.

[6] B. Brown and E. Laurier, "The normal natural troubles of driving with gps," in *ACM SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI 2012. Austin, TX, USA: ACM, May 2012.

 ISBN: 978-1-61208-203-5

Differentially Amplitude- and Phase-Encoded QAM in Amplify-and-Forward Multiple Relay System Over Nakagami-m Fading Channels

Chi-Hua Huang and Char-Dir Chung
Graduate Institute of Communication Engineering
National Taiwan University
Taipei 10617, Taiwan
f95942046@ntu.edu.tw and cdchung@ntu.edu.tw

***Abstract*— This paper studies the differentially amplitude- and phase-encoded (DAPE) quadrature amplitude modulation (QAM) for the amplify-and-forward multiple relay system over independent Nakagami-m fading links. A simple equal gain combining (EGC) receiver is proposed to noncoherently combine received signals from direct and multiple relay links and then detect the DAPE QAM signals without any link side information. Based on Beaulieu's series approach, an efficient bit error probability (BEP) upper bound computation formula is derived for the EGC receiver. Performance results show that the EGC receiver for DAPE QAM provides better BEP than the conventional receiver for differential phase shift keying modulation with the same constellation size.**

***Keywords*—Amplify-and-forward relay, differential detection, Nakagami-m fading, star-QAM.**

I. INTRODUCTION

Cooperative relaying technique [1]-[5] is an efficacious method to realize distributed spatial diversity through the cooperation of relay nodes in terms of radio resource and signal processing. In [1]-[2], various cooperative transmission protocols based on half-duplex mode were developed to achieve cooperative diversity. As mentioned therein, the amplify-and-forward (AF) relaying protocol is a simple relaying scheme in which the relays amplify the received signals from the source and retransmit them to the destination. This paper only considers the AF protocol due to its simpler operation and lower complexity required at relays.

Coherent detection is considered in most relay systems, which is based on the assumption that the destination can obtain the channel impulse response (CIR) characteristics of all transmission links [1]-[2]. However, in fast fading environments, it is difficult to obtain the accurate CIR information at the destination through delicate link estimation. To reduce the overhead for complicated link estimation in relay systems, differential modulation has been investigated in [3]-[5]. For the AF multiple relay systems using differential phase shift keying (DPSK) modulation, a weighted gain combining (WGC) receiver was developed in conjunction with approximate bit error probability (BEP) analysis over independent Rayleigh [3] and Nakagami-m [4] fading links. Optimal power allocation for AF multiple relay system based on a simple BEP upper bound was also considered in [3], [5] to improve the overall system performance. Despite a wealth of past studies, the DPSK modulation decreases dramatically in power efficiency when the modulation alphabet size is increased.

The differentially amplitude- and phase-encoded (DAPE) (J_a, J_p) quadrature amplitude modulation (QAM) [6]-[7] is an effective technique to achieve high bit rate transmission without CIR information at the receiver side. DAPE (J_a, J_p) QAM employs a star constellation with J_a concentric amplitude rings, each containing J_p phasors, and sequentially encodes information onto the level changes of amplitude and phase between currently and previously transmitted signals. Due to its higher spectral efficiency and better BEP performance than DPSK, DAPE QAM was extensively studied in [6]-[7] for single link communications. However, DAPE QAM is not yet investigated for relay communications.

This paper studies the DAPE (J_a, J_p) QAM signals for AF multiple relay system over independent Nakagami-m fading links. Section II depicts the system and channel models. In Section III, the conventional EGC receiver in [6]-[7] is adopted to cater to the AF multiple relay system and shown to operate without CIR or any sort of channel state information (CSI) when compared to the noncoherent DPSK WGC and coherent QAM receivers in [1]-[5]. Inevitably, in order to meet the corresponding relay power constraint, the knowledge of average source-relay signal-to-noise power ratio (SNR) is still required at each relay for the EGC receiver, which is a design requirement typical in all differentially coherent AF multiple relay systems [3]-[5]. Based on a union bound argument and a convergent infinite series approach [8], the analytical upper bound of BEP is derived for the EGC receiver over independent Nakagami-m fading links in Section IV. Performance results in Section V show that the EGC receiver for DAPE QAM provides better BEP than the WGC receiver for DPSK with the same constellation size.

Nomenclature: $\mathbb{E}\{\cdot\}$ is the expectation. $\mathcal{Z}_K$ and $\mathcal{Z}_K^+$ denote the integer sets $\{0, 1, ..., K-1\}$ and $\{1, 2, ..., K\}$, respectively. $\mathrm{Re}\{x\}$ is the real part of a complex number x and $\mathrm{Im}\{x\}$ the imaginary part. m mod M denotes the modulo-M value of integer m. Superscript $*$ is the complex conjugate.

 ISBN: 978-1-61208-203-5

II. System and Channel Models

This paper considers the DAPE (J_a, J_p) QAM [6] as the modulation approach. In the DAPE (J_a, J_p) QAM scheme, the transmitted bit sequence corresponding to a nominal N-symbol block is first grouped into $N-1$ independent symbol pairs $\{(\Delta a_n, \Delta b_n)\}_{n=1}^{N-1}$ where $\Delta a_n \in \mathcal{Z}_{J_a}$ and $\Delta b_n \in \mathcal{Z}_{J_p}$. Here, all $J_a J_p$ possible pairs for each $(\Delta a_n, \Delta b_n)$ are assumed to be transmitted equally likely and the binary Gray labeling is used to denote Δa_n and Δb_n individually. The symbols Δa_n and Δb_n are then used to determine the amplitude ratio and the phase difference, respectively, between two consecutively transmitted symbol signals. With $\mu > 1$, the nth transmitted symbol signal is given by $x_n = \lambda \mu^{a_n} e^{j2\pi b_n / J_p}$ for $n \in \mathcal{Z}_{N-1}^+$ and $x_0 = \lambda$ where $\lambda = \sqrt{J_a(\mu^2-1)/(\mu^{2J_a}-1)}$ is used to normalize the signal constellation to unit energy. a_n and b_n represent respectively the amplitude and phase levels of x_n for $n \in \mathcal{Z}_{N-1}^+$ and are given by $a_n = (a_{n-1} + \Delta a_n) \bmod J_a$ and $b_n = (b_{n-1} + \Delta b_n) \bmod J_p$, with $a_0 = b_0 = 0$. Notably, DAPE QAM with $J_a = 1$ corresponds to J_p-ary DPSK.

Consider a multiple relay system which consists of a source node s, a destination node d, and L relay nodes $\{r_l\}_{l=1}^L$. All relays are in a half-duplex mode which can not transmit and receive both in time and frequency simultaneously. In order to avoid mutual interference, it is assumed that the signal transmission scheme includes $L+1$ distinct phases in time. In the first phase, the source broadcasts the signals $\{x_n\}_{n=0}^{N-1}$ with power P_s to all relays and destination. The corresponding nth received symbol signals at destination and relay r_l are

$$y_{sd,n} = \sqrt{P_s \Upsilon_{sd}} h_{sd} x_n + w_{sd,n} \quad (1)$$

$$y_{sl,n} = \sqrt{P_s \Upsilon_{sl}} h_{sl} x_n + w_{sl,n},\ l \in \mathcal{Z}_L^+ \quad (2)$$

for $n \in \mathcal{Z}_N$. In the $(l+1)$th phase, the relay r_l amplifies its received symbol signals at the first phase to produce $u_{l,n} = \alpha_l y_{sl,n}$ where α_l is the amplification factor and retransmits $\{u_{l,n}\}_{n=0}^{N-1}$ with power P_l to the destination. At the destination, the nth received symbol signal corresponding to the $(l+1)$th phase is

$$y_{ld,n} = \sqrt{P_l \Upsilon_{ld}} h_{ld} u_{l,n} + w_{ld,n}. \quad (3)$$

In the $(L+1)$th phase, the destination then combines all the signals received in all the phases to make a final decision on the information carried by the N-symbol block.

In the above modeling, the coefficients Υ_{ij} and h_{ij} denote respectively the path loss which depends on the geographical distribution of the relay network and the CIR between node i and node j for $ij \in \{sd, \{sl\}_{l=1}^L, \{ld\}_{l=1}^L\}$. Each h_{ij} is assumed constant over an N-symbol block and varies block by block. For analytical convenience, h_{ij}'s are modeled to be independent and have Nakagami-m fading amplitude. For Nakagami-m fading, h_{ij} has independent Nakagami-m distributed amplitude and uniformly distributed phase. The probability density function of a Nakagami-m amplitude $z_{ij} \triangleq |h_{ij}|$ with normalized average power $\mathbb{E}\{|h_{ij}|^2\} = 1$ is $f_{z_{ij}}(z_{ij}) = \frac{2}{\Lambda(m_{ij})} m_{ij}^{m_{ij}} z_{ij}^{2m_{ij}-1} \exp\{-m_{ij} z_{ij}^2\}$ where m_{ij} is the fading parameter, defined by $m_{ij} \triangleq 1/\mathbb{E}\{(z_{ij}^2 - 1)^2\} \geq 1/2$ [9], and $\Lambda(m) = \int_0^\infty e^{-t} t^{m-1} dt$ is the Gamma function [10]. The additive white Gaussian noise (AWGN) samples $\{w_{ij,n} | ij \in \{sd, \{sl\}_{l=1}^L, \{ld\}_{l=1}^L\}, n \in \mathcal{Z}_N\}$ are modeled as independent and identically distributed (iid) circularly symmetric complex Gaussian random variables (CGRVs) with mean 0 and identical variance σ_w^2. Furthermore, AWGN and fading gains are assumed independent.

In the following, $\gamma_{ij} \triangleq P_i \Upsilon_{ij} / \sigma_w^2$ denotes the average link SNR per symbol from node i to node j for $ij \in \{sd, \{sl\}_{l=1}^L, \{ld\}_{l=1}^L\}$. It is assumed that the total transmission power $P_s + \sum_{l=1}^L P_l = P_t$ is fixed. In the case, $\gamma_t \triangleq P_t/\sigma_w^2$ and $\gamma_b \triangleq \gamma_t / \log_2(J_a J_p)$ are respectively the total transmitted SNRs per symbol and per bit. Moreover, $\widehat{\widehat{\gamma}}_{sl} \triangleq (1 + \widehat{\widehat{\delta}}_{sl})\gamma_{sl}$ for $l \in \mathcal{Z}_L^+$ and $\widehat{\gamma}_{ij} \triangleq (1 + \widehat{\delta}_{ij})\gamma_{ij}$ for $ij \in \{sd, \{sl\}_{l=1}^L, \{ld\}_{l=1}^L\}$ denote respectively the estimate of average link SNR γ_{sl} made at relay r_l and the estimate of average link SNR γ_{ij} made at destination where $\widehat{\widehat{\delta}}_{sl}$ and $\widehat{\delta}_{ij}$ represent the normalized SNR estimation errors. At relay r_l, the amplification factor α_l has to satisfy the respective average relay transmission power constraint $\mathbb{E}\{\left|\sqrt{P_l}\alpha_l y_{sl,n}\right|^2\} = P_l$ and is given by $\alpha_l = 1/(\sigma_w \sqrt{\gamma_{sl}+1})$ for $l \in \mathcal{Z}_L^+$ [3]-[5]. Notably, the average link SNR γ_{sl} is required for realizing α_l and can be measured through conventional SNR estimation methods [11]. With SNR estimation, γ_{sl} in α_l can be replaced by an estimate $\widehat{\widehat{\gamma}}_{sl}$ for capturing the impact of SNR estimation error. Since the noise variance σ_w^2 remains constant over long periods of time in practice, it is assumed to be perfectly measured.

III. Decision Algorithm

In this section, a symbol-by-symbol decision algorithm for detecting $(\Delta a_n, \Delta b_n)$ is derived based on the received symbol signals $\{y_{sd,n-1}, y_{sd,n}\}$ and $\{y_{ld,n-1}, y_{ld,n}\}_{l=1}^L$ for $n \in \mathcal{Z}_{N-1}^+$. For notational brevity, the subscript n in $(\Delta a_n, \Delta b_n)$ is dropped below.

The EGC receiver [6]-[7] is commonly used for demodulating DAPE (J_a, J_p) QAM with multiple received antennas and can be applied here for the AF multiple relay system. Specifically, the EGC receiver makes amplitude and phase decisions separately. The amplitude decision is based on detecting the amplitude ratio of successively received symbol signals in $\{y_{sd,n-1}, y_{sd,n}\}$ and $\{y_{ld,n-1}, y_{ld,n}\}_{l=1}^L$ in conjunction with square-law combining. The test metric on amplitude ratio is given by

$$W_a = \frac{|y_{sd,n}|^2 + \sum_{l=1}^L q_l |y_{ld,n}|^2}{|y_{sd,n-1}|^2 + \sum_{l=1}^L q_l |y_{ld,n-1}|^2} \quad (4)$$

where $q_l \triangleq 1$ for $l \in \mathcal{Z}_L^+$. Thus, the amplitude decision rule is to declare $\Delta\widehat{a} = k$ if $W_a \in R_k$, where the amplitude decision region R_k is defined by $R_k \triangleq \{W_a | \eta_k^2 \leq W_a < \eta_{k+1}^2$ or $\eta_{k-J_a}^2 \leq W_a < \eta_{k-J_a+1}^2\}$ for $k \in \mathcal{Z}_{J_a-1}^+$ and $R_0 \triangleq \{W_a | \eta_0^2 \leq W_a < \eta_1^2\}$ with $0 = \eta_{-J_a+1} < \eta_{-J_a+2} < \ldots < \eta_{J_a} = \infty$. The amplitude decision threshold η_k is given by

 ISBN: 978-1-61208-203-5

$\eta_k = \mu^{k-1/2}$ for $k = -J_a + 2, -J_a + 3, ..., J_a - 1$ as in [6].[1] On the other hand, the differential phase decision is based on the conventional product detector for demodulating J_p-ary DPSK with multiple observations [9], and the test metric on phase difference is given by

$$W_p = y_{sd,n} y^*_{sd,n-1} + \sum_{l=1}^{L} q_l y_{ld,n} y^*_{ld,n-1}. \quad (5)$$

Thus, the phase decision rule is to declare $\Delta\widehat{b}$ if $\mathrm{Re}\{W_p e^{-j2\pi\Delta b/J_p}\}$ is maximized when $\Delta b = \Delta\widehat{b}$.

Notably, for the WGC receiver detecting J_p-ary DPSK in [3]-[5], the test metric on phase difference is the same as (5) with $q_l = 1$ replaced by $q_l = (1+\widehat{\gamma}_{sl})/(1+\widehat{\gamma}_{sl}+\widehat{\gamma}_{ld})$ for $l \in \mathcal{Z}_L^+$, where the link SNR estimates $\widehat{\gamma}_{sl}$ and $\widehat{\gamma}_{ld}$ can be measured prior to data detection at destination through SNR estimation methods [11].[2] Therefore, the phase decision rule of the WGC receiver for J_p-ary DPSK is to declare $\Delta\widehat{b}$ if $\mathrm{Re}\{W_p e^{-j2\pi\Delta b/J_p}\}$ is maximized when $\Delta b = \Delta\widehat{b}$.

IV. BEP Analysis

In this section, the BEP upper bound of EGC receiver is analyzed below for independent Nakagami-m fading links.

A) BEP Characteristics: The average BEP can be generally expressed as [6]-[7]

$$\mathcal{P}_b = \frac{\mathcal{P}_a \log_2 J_a + \mathcal{P}_p \log_2 J_p}{\log_2(J_a J_p)} \quad (6)$$

where $\mathcal{P}_a$ and $\mathcal{P}_p$ are the average BEPs for detecting amplitude ratio and phase difference, respectively. Specifically, $\mathcal{P}_a$ and the union bound for $\mathcal{P}_p$ are given by

$$\mathcal{P}_a = \frac{1}{J_a} \sum_{\substack{\Delta a, \Delta\widehat{a}=0 \\ \Delta a \neq \Delta\widehat{a}}}^{J_a-1} \frac{c(\Delta a, \Delta\widehat{a})}{\log_2 J_a} \mathcal{P}_1(\Delta\widehat{a}|\Delta a) \quad (7)$$

$$\mathcal{P}_p \leq \frac{1}{J_p} \sum_{\substack{\Delta b, \Delta\widehat{b}=0 \\ \Delta b \neq \Delta\widehat{b}}}^{J_p-1} \frac{c(\Delta b, \Delta\widehat{b})}{\log_2 J_p} \cdot \frac{1}{J_a^2} \sum_{a_{n-1}=0}^{J_a-1} \sum_{a_n=0}^{J_a-1} \mathcal{P}_2(\Delta\widehat{b}|a_{n-1}, a_n, \Delta b). \quad (8)$$

Here, $c(i,j)$ is the Hamming distance between the binary representations of i and j. $\mathcal{P}_1(\Delta\widehat{a}|\Delta a)$ represents the probability of deciding $\Delta\widehat{a}$ (i.e., $W_a \in R_{\Delta\widehat{a}}$) given that Δa is transmitted. $\mathcal{P}_2(\Delta\widehat{b}|a_{n-1}, a_n, \Delta b) \triangleq \Pr\{\mathrm{Re}\{W_p e^{-j2\pi\Delta b/J_p}\} < \mathrm{Re}\{W_p e^{-j2\pi\Delta\widehat{b}/J_p}\}|a_{n-1}, a_n, \Delta b\}$ denotes the pairwise error probability that $\mathrm{Re}\{W_p e^{-j2\pi\Delta\widehat{b}/J_p}\}$ is larger than $\mathrm{Re}\{W_p e^{-j2\pi\Delta b/J_p}\}$ given that a_{n-1}, a_n, and Δb are transmitted. Using (7) and (8) in (6) gives an upper bound to the BEP of the EGC receiver.[3]

The evaluation of $\mathcal{P}_1(\Delta\widehat{a}|\Delta a)$ has to be separately treated for two cases $\Delta\widehat{a} = 0$ and $\Delta\widehat{a} \neq 0$ since different decision region formats are involved. $\mathcal{P}_1(\Delta\widehat{a}|\Delta a)$ is given by[4]

$$\mathcal{P}_1(0|\Delta a) = \frac{1}{J_a} \sum_{a_{n-1}=0}^{J_a-1-\Delta a} \mathcal{P}_3(0|a_{n-1}, a_{n-1}+\Delta a) + \frac{1}{J_a} \sum_{a_{n-1}=J_a-\Delta a}^{J_a-1} \mathcal{P}_3(0|a_{n-1}, a_{n-1}+\Delta a - J_a) \quad (9)$$

when $\Delta\widehat{a} = 0$, and

$$\mathcal{P}_1(\Delta\widehat{a}|\Delta a) = \frac{1}{J_a} \sum_{a_{n-1}=0}^{J_a-1-\Delta a} [\mathcal{P}_3(\Delta\widehat{a}|a_{n-1}, a_{n-1}+\Delta a) + \mathcal{P}_3(\Delta\widehat{a} - J_a|a_{n-1}, a_{n-1}+\Delta a)] + \frac{1}{J_a} \sum_{a_{n-1}=J_a-\Delta a}^{J_a-1} [\mathcal{P}_3(\Delta\widehat{a}|a_{n-1}, a_{n-1}+\Delta a - J_a) + \mathcal{P}_3(\Delta\widehat{a} - J_a|a_{n-1}, a_{n-1}+\Delta a - J_a)] \quad (10)$$

when $\Delta\widehat{a} \in \mathcal{Z}_{J_a-1}^+$. In (9) and (10), $\mathcal{P}_3(k|a_{n-1}, a_n) \triangleq \Pr\{\eta_k^2 \leq W_a < \eta_{k+1}^2|a_{n-1}, a_n\}$ denotes the conditional probability of event $\{\eta_k^2 \leq W_a < \eta_{k+1}^2\}$ given that a_{n-1} and a_n are transmitted. Note here that $\Pr\{\eta_k^2 \leq W_a < \eta_{k+1}^2|a_{n-1}, a_n\}$ can be alternatively expressed and thus evaluated as $\Pr\{W_a < \eta_{k+1}^2|a_{n-1}, a_n\} - \Pr\{W_a < \eta_k^2|a_{n-1}, a_n\}$.

Both events $\{W_a < \eta_k^2\}$ and $\{\mathrm{Re}\{W_p e^{-j2\pi\Delta b/J_p}\} < \mathrm{Re}\{W_p e^{-j2\pi\Delta\widehat{b}/J_p}\}\}$ in evaluating respectively $\Pr\{W_a < \eta_k^2|a_{n-1}, a_n\}$, which is required for computing $\mathcal{P}_1(\Delta\widehat{a}|\Delta a)$ in (7), and $\mathcal{P}_2(\Delta\widehat{b}|a_{n-1}, a_n, \Delta b)$ in (8) can be conveniently unified in the form of $\{X_i + \sum_{l=1}^{L} Y_{l,i} < 0\}$, in which the variables X_i and $Y_{l,i}$ are defined as $X_i \triangleq A_i |y_{sd,n}/\sigma_w|^2 + B_i |y_{sd,n-1}/\sigma_w|^2 + 2\,\mathrm{Re}\{C_i y_{sd,n} y^*_{sd,n-1}/\sigma_w^2\}$ and $Y_{l,i} \triangleq D_{l,i} |y_{ld,n}/\sigma_w|^2 + E_{l,i} |y_{ld,n-1}/\sigma_w|^2 + 2\,\mathrm{Re}\{F_{l,i} y_{ld,n} y^*_{ld,n-1}/\sigma_w^2\}$, respectively, for $l \in \mathcal{Z}_L^+$ and $i \in \{a, p\}$.[5] The coefficient vector $(A_i, B_i, C_i, D_{l,i}, E_{l,i}, F_{l,i})$ is set to $(1, -\eta_k^2, 0, q_l, -q_l\eta_k^2, 0)$ for event $\{W_a < \eta_k^2\}$ (when $i = a$) and $(0, 0, e^{-j2\pi\Delta b/J_p} - e^{-j2\pi\Delta\widehat{b}/J_p}, 0, 0, q_l(e^{-j2\pi\Delta b/J_p} - e^{-j2\pi\Delta\widehat{b}/J_p}))$ for event $\{\mathrm{Re}\{W_p e^{-j2\pi\Delta b/J_p}\} < \mathrm{Re}\{W_p e^{-j2\pi\Delta\widehat{b}/J_p}\}\}$ (when $i = p$). In terms of the unified format, $\Pr\{W_a < \eta_k^2|a_{n-1}, a_n\}$ and $\mathcal{P}_2(\Delta\widehat{b}|a_{n-1}, a_n, \Delta b)$ can be both expressed as $\Pr\{X_i + \sum_{l=1}^{L} Y_{l,i} < 0|\mathcal{S}\}$ given $\mathcal{S} \triangleq \{a_{n-1}, a_n, \Delta b\}$.

[1] Due to cyclic differential amplitude encoding, when $\Delta a \neq 0$ the two possible cases $|x_n|/|x_{n-1}| > 1$ and $|x_n|/|x_{n-1}| < 1$ must be considered in amplitude decision. This explains that R_k for $k \neq 0$ consists of two disjoint regions where $\{W_a|\eta_k^2 \leq W_a < \eta_{k+1}^2\}$ accounts for the case $|x_n| > |x_{n-1}|$ and $\{W_a|\eta_{k-J_a}^2 \leq W_a < \eta_{k-J_a+1}^2\}$ for the case $|x_n| < |x_{n-1}|$.

[2] As indicated, the EGC receiver for DAPE QAM operates without any CSI, while the knowledge of average SNR estimates on source-relay and relay-destination links is required for realizing the WGC receiver for DPSK.

[3] As mentioned previously, the test metric on phase difference W_p is of the same form for both EGC and WGC receivers except for gain variables $\{q_l\}_{l=1}^{L}$. Thus, when $J_a = 1$, (6) also gives an upper bound to the BEP of the WGC receiver for J_p-ary DPSK [3]-[5].

[4] By default, $\sum_{k=n}^{m} = 0$, if $n > m$.

[5] Here, the subscript i used for the variables X_i and $Y_{l,i}$ denotes the respective cases for event $\{W_a < \eta_k^2\}$ (when $i = a$) and event $\{\mathrm{Re}\{W_p e^{-j2\pi\Delta b/J_p}\} < \mathrm{Re}\{W_p e^{-j2\pi\Delta\widehat{b}/J_p}\}\}$ (when $i = p$).

Based on Beaulieu's convergent infinite series approach proposed by [8], $\Pr\{X_i+\sum_{l=1}^{L} Y_{l,i} < 0|\mathcal{S}\}$ can be efficiently evaluated within a predetermined accuracy as

$$\Pr\left\{X_i + \sum_{l=1}^{L} Y_{l,i} < 0|\mathcal{S}\right\} \approx \frac{1}{2} - \frac{2}{\pi} \sum_{\substack{m=1 \\ m\ odd}}^{\infty} \frac{1}{m} \cdot \mathrm{Im}\left\{\Phi_{X_i}(jm\omega_0|\mathcal{S}) \prod_{l=1}^{L} \Phi_{Y_{l,i}}(jm\omega_0|\mathcal{S})\right\} \quad (11)$$

where $j \triangleq \sqrt{-1}$, $\omega_0 \triangleq 2\pi/T$ with T being the period of the square wave used in deriving the series, and $\Phi_{X_i}(j\omega|\mathcal{S})$ and $\Phi_{Y_{l,i}}(j\omega|\mathcal{S})$ represent respectively the conditional characteristic functions (CFs) of X_i and $Y_{l,i}$ given $\mathcal{S}$. The series in (11) converges pointwise to the true distribution value within a specific accuracy when T is chosen large enough and can be thus truncated to a desired accuracy. To facilitate the evaluation of (11), analytical expressions for $\Phi_{X_i}(j\omega|\mathcal{S})$ and $\Phi_{Y_{l,i}}(j\omega|\mathcal{S})$ are derived in the following.

B) Characteristic Functions $\Phi_{X_i}(j\omega|\mathcal{S})$ and $\Phi_{Y_{l,i}}(j\omega|\mathcal{S})$: Conditioned on $\mathcal{S}$ and h_{sd}, $y_{sd,n}$ and $y_{sd,n-1}$ are jointly CGRVs, and thus X_i is a conditionally Gaussian quadratic sum (GQS) [9]. Quoting [9, eq. (B-5)], $\Phi_{X_i}(j\omega|\mathcal{S}, h_{sd})$ is readily given by

$$\Phi_{X_i}(j\omega|\mathcal{S},z_{sd}) = \frac{1}{H_i(\omega)} \exp\left\{\frac{G_i(\omega) z_{sd}^2}{H_i(\omega)}\right\} \quad (12)$$

where $G_i(\omega) \triangleq \gamma_{sd}\{\omega^2(A_iB_i - |C_i|^2)(|x_n|^2 + |x_{n-1}|^2) + j\omega[A_i|x_n|^2 + B_i|x_{n-1}|^2 + 2\,\mathrm{Re}\{C_i x_n x_{n-1}^*\}]\}$ and $H_i(\omega) \triangleq 1 - j\omega(A_i + B_i) + \omega^2(|C_i|^2 - A_iB_i)$. Similarly, conditioned on $\mathcal{S}$, h_{sl}, and h_{ld}, $y_{ld,n}$ and $y_{ld,n-1}$ are jointly CGRVs, and thus $Y_{l,i}$ is also a conditionally GQS. $\Phi_{Y_{l,i}}(j\omega|\mathcal{S}, z_{sl}, z_{ld})$ can also be derived from [9, eq. (B-5)] as

$$\Phi_{Y_{l,i}}(j\omega|\mathcal{S}, z_{sl}, z_{ld}) = \frac{1}{V_{l,i}(\omega|z_{ld})} \exp\left\{\frac{U_{l,i}(\omega|z_{ld}) z_{sl}^2}{V_{l,i}(\omega|z_{ld})}\right\} \quad (13)$$

where $U_{l,i}(\omega|z_{ld}) \triangleq \gamma_{sl}\upsilon\{\omega^2\kappa(1+\upsilon)(|x_n|^2 + |x_{n-1}|^2) + j\omega[D_{l,i}|x_n|^2 + E_{l,i}|x_{n-1}|^2 + 2\,\mathrm{Re}\{F_{l,i}x_n x_{n-1}^*\}]\}$ and $V_{l,i}(\omega|z_{ld}) \triangleq 1 - (1+\upsilon)^2[j\omega(D_{l,i} + E_{l,i})/(1+\upsilon) + \omega^2\kappa]$ with $\upsilon \triangleq \gamma_{ld} z_{ld}^2 \alpha_l^2 \sigma_w^2$ and $\kappa \triangleq D_{l,i}E_{l,i} - |F_{l,i}|^2$. Thus, $\Phi_{X_i}(j\omega|\mathcal{S})$ and $\Phi_{Y_{l,i}}(j\omega|\mathcal{S})$ can be respectively obtained by averaging $\Phi_{X_i}(j\omega|\mathcal{S},z_{sd})$ over the density of z_{sd} and $\Phi_{Y_{l,i}}(j\omega|\mathcal{S}, z_{sl}, z_{ld})$ over the joint density of z_{sl} and z_{ld}, as given below.

1) Characteristic Function $\Phi_{X_i}(j\omega|\mathcal{S})$: Using [12, eq. (3.478.1)], averaging $\Phi_{X_i}(j\omega|\mathcal{S}, z_{sd})$ over $f_{z_{sd}}(z_{sd})$ gives the CF $\Phi_{X_i}(j\omega|\mathcal{S})$ as

$$\Phi_{X_i}(j\omega|\mathcal{S}) = \frac{1}{H_i(j\omega)} \left[\frac{H_i(j\omega)}{H_i(j\omega) - G_i(j\omega)/m_{sd}}\right]^{m_{sd}}. \quad (14)$$

2) Characteristic Function $\Phi_{Y_{l,i}}(j\omega|\mathcal{S})$: Note that $\Phi_{Y_{l,i}}(j\omega|\mathcal{S}, z_{sl}, z_{ld})$ in (13) is exactly of the same expression as $\Phi_{X_i}(j\omega|\mathcal{S},z_{sd})$ in (12) with z_{sd}, $G_i(\omega)$, and $H_i(\omega)$ respectively replaced by z_{sl}, $U_{l,i}(\omega|z_{ld})$, and $V_{l,i}(\omega|z_{ld})$. Thus, by virtue of the independence between z_{sl} and z_{ld}, $\Phi_{Y_{l,i}}(j\omega|\mathcal{S}, z_{ld})$ can be similarly obtained as the expressions in (14) with $G_i(\omega) \rightarrow U_{l,i}(\omega|z_{ld})$, $H_i(\omega) \rightarrow V_{l,i}(\omega|z_{ld})$, and $m_{sd} \rightarrow m_{sl}$ after averaging $\Phi_{Y_{l,i}}(j\omega|\mathcal{S}, z_{sl}, z_{ld})$ over $f_{z_{sl}}(z_{sl})$. By averaging the resultant $\Phi_{Y_{l,i}}(j\omega|\mathcal{S}, z_{ld})$ over $f_{z_{ld}}(z_{ld})$, the CF $\Phi_{Y_{l,i}}(j\omega|\mathcal{S})$ can be obtained as

$$\Phi_{Y_{l,i}}(j\omega|\mathcal{S}) = \int_0^{\infty} \Phi_{Y_{l,i}}(j\omega|\mathcal{S}, z_{ld}) f_{z_{ld}}(z_{ld}) dz_{ld}. \quad (15)$$

Unfortunately, the single integral in (15) is difficult, if not impossible, to complete. However, it can be numerically computed by Gaussian quadrature rule [10, eq. (25.4.45)].

V. Performance Results

This section illustrates the BEP results of the EGC receiver in conjunction with DAPE (J_a, J_p) QAM formats with constellation sizes $J_aJ_p = 64$ over iid Nakagami-m fading links. For notational brevity, the subscript ij in Nakagami-m fading parameter m_{ij} is dropped. The BEP characteristics for the EGC receiver is evaluated by using the BEP upper bound expressions (6)-(8) as well as Monte Carlo simulation. For simplicity, it is also assumed that all source-relay path losses, all relay-destination path losses, and all transmission powers are respectively identical with $\Upsilon_{sl} = \Upsilon_{sr}$, $\Upsilon_{ld} = \Upsilon_{rd}$, and $P_l = P_s = P_t/(L+1)$ for all $l \in \mathcal{Z}_L^+$. In the presence of SNR estimation errors, it is further assumed that all source-relay links suffer the same level of error and so do relay-destination links, with $\widehat{\widehat{\delta}}_{sl} = \widehat{\widehat{\delta}}_{sr}$, $\widehat{\delta}_{sl} = \widehat{\delta}_{sr}$, and $\widehat{\delta}_{ld} = \widehat{\delta}_{rd}$ for all $l \in \mathcal{Z}_L^+$. Moreover, the estimate $\widehat{\widehat{\gamma}}_{sl}$ made at relay r_l is assumed to be delivered to the destination through a very reliable transmission link, and thus $\widehat{\widehat{\gamma}}_{sl}$ is equal to the estimate $\widehat{\gamma}_{sl}$ made at destination for all $l \in \mathcal{Z}_L^+$ (i.e., $\widehat{\widehat{\delta}}_{sr} = \widehat{\delta}_{sr}$).

The BEP of a DAPE QAM system with a fixed constellation size depends on the setting of ring ratio μ and (J_a, J_p) [6]-[7]. By minimizing the BEP upper bound expressions, the optimization of μ and (J_a, J_p) can be achieved. As observed by the authors, when μ is optimized, DAPE $(4, 16)$ QAM gives the best BEP performance in all possible 64-point constellations for wide ranges of m and γ_b. Moreover, the BEP upper bound achieves nearly the best when μ is fixed to 1.4 for DAPE $(4, 16)$ QAM, and thus this value for μ is adopted below.

Figs. 1-2 compare the BEP characteristics among the EGC receiver for DAPE $(4, 16)$ QAM and 64-ary DPSK and the WGC receiver for 64-ary DPSK [3]-[5]. Fig. 1 illustrates the BEP characteristics for various scenarios with different geographical relay locations. The scenarios represent that all relays are close to the source when Υ_{sr} is larger than Υ_{rd} and that all relays are close to the destination when Υ_{rd} is larger than Υ_{sr} [3]. As indicated, the EGC receiver under the scenario $\Upsilon_{sr} > \Upsilon_{rd}$ performs better than that under the scenario $\Upsilon_{rd} > \Upsilon_{sr}$, but this performance prevalence is reversed for the WGC receiver. The difference in the performance trends between the EGC and WGC receivers comes from the settings of the gain variables $\{q_l\}_{l=1}^{L}$ that the average source-relay and relay-destination link SNRs are not leveraged by the

 ISBN: 978-1-61208-203-5

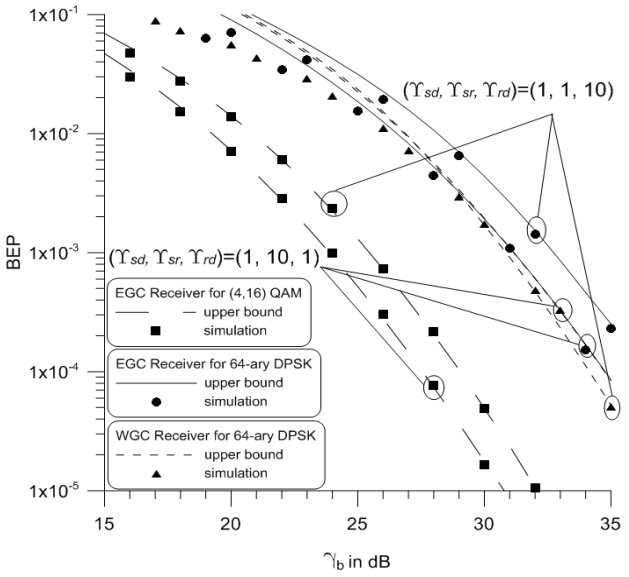


Fig. 1. BEP versus SNR/bit characteristics of EGC receiver for DAPE $(4, 16)$ QAM and 64-ary DPSK and WGC receiver for 64-ary DPSK with $\widehat{\delta}_{sr} = \widehat{\delta}_{rd} = 0$, $m = 1.5$, and $L = 2$.

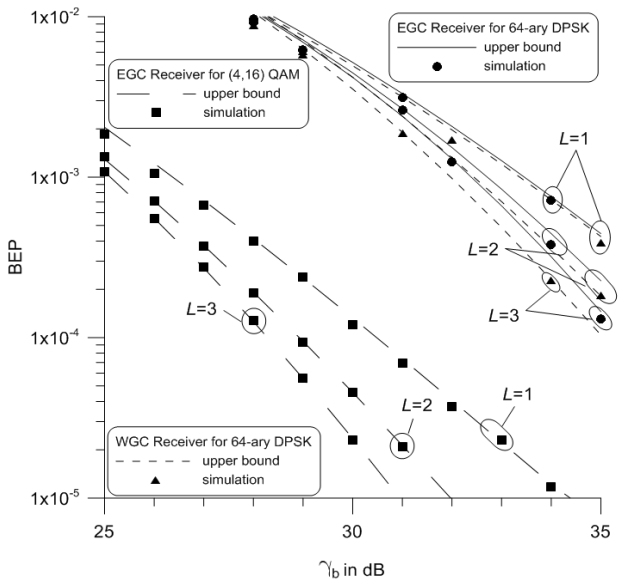


Fig. 2. BEP versus SNR/bit characteristics of EGC receiver for DAPE $(4, 16)$ QAM and 64-ary DPSK and WGC receiver for 64-ary DPSK with $\widehat{\delta}_{sr} = \widehat{\delta}_{rd} = 0$, $m = 1.5$, and $(\Upsilon_{sd}, \Upsilon_{sr}, \Upsilon_{rd}) = (1, 1, 1)$.

EGC receiver but used by the WGC receiver. Thus, the EGC receiver suffers the BEP performance degradation under the scenario $\Upsilon_{rd} > \Upsilon_{sr}$. As also observed, when $\Upsilon_{sr} \gg \Upsilon_{rd}$ (e.g., $(\rho_{sr}, \rho_{rd}) = (10, 1)$), the EGC and WGC receivers for DPSK perform almost the same because the gain variable of the WGC receiver $q_l \approx 1$ for all $l \in \mathcal{Z}_L$, and both EGC and WGC receivers for DPSK are approximately equivalent. Fig. 2 shows that the diversity reception can effectively improve the BEP performance. This performance improvement is, however, achieved at the cost of using more relays as well as reducing overall network throughput [4]. As indicated in Figs. 1-2, the EGC receiver for DAPE QAM significantly outperforms the EGC and WGC receivers for DPSK. As also observed, the BEP upper bounds for both EGC and WGC receivers are in the better agreement with simulation when the average link SNRs are sufficiently larger and $L > 1$.

Fig. 3 demonstrates the BEP characteristics of EGC receiver for DAPE QAM and DPSK as well as WGC receiver for DPSK with link SNR estimation errors. Because the test metrics depend on link SNR estimates for the WGC receiver but not for the EGC receiver, in addition to the requirement that all relay gains depend on source-relay link SNR estimates, the EGC receiver for DAPE QAM degrades in BEP less than the WGC receiver in the presence of link SNR estimation errors. This explains that the EGC receiver is less sensitive to link SNR estimation errors than the WGC receiver.

Fig. 4 compares the BEP characteristics among the EGC receiver for DAPE $(4, 16)$ QAM, the WGC receiver for 64-ary DPSK, and maximum-likelihood (ML) receiver [2] for coherent 64-ary rectangular QAM with Gray labeling. For the coherent ML receiver, the decision rule is to declare the

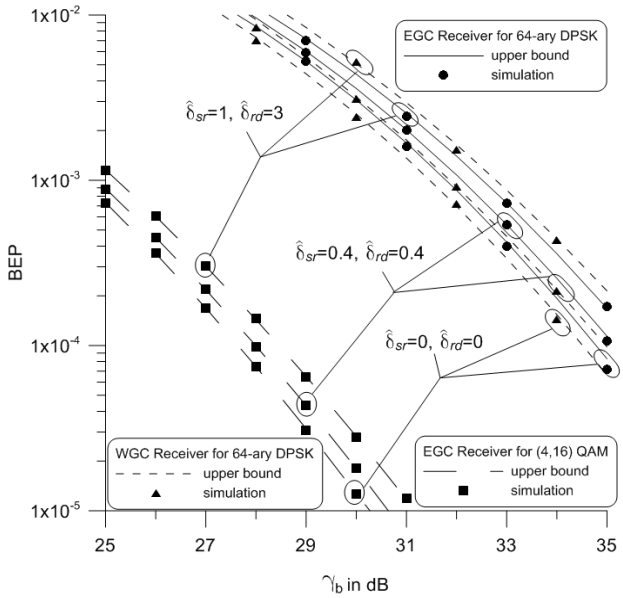


Fig. 3. BEP versus SNR/bit characteristics of EGC receiver for DAPE $(4, 16)$ QAM and 64-ary DPSK and WGC receiver for 64-ary DPSK with $m = 2$, $L = 2$, and $(\Upsilon_{sd}, \Upsilon_{sr}, \Upsilon_{rd}) = (1, 1, 1)$.

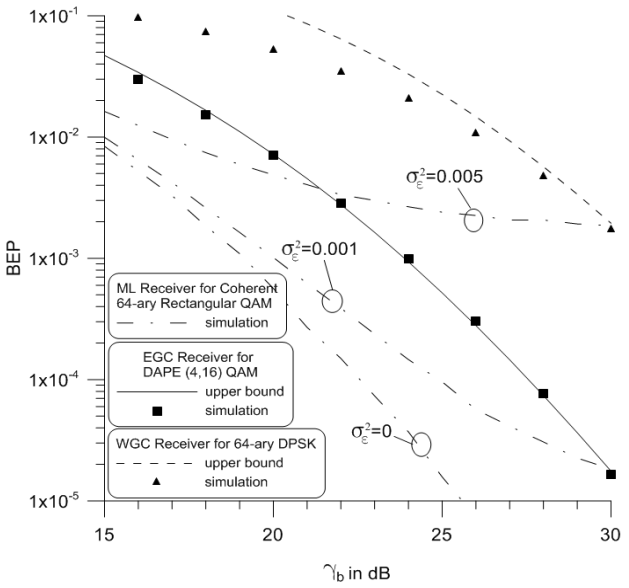


Fig. 4. BEP versus SNR/bit characteristics of EGC receiver for DAPE $(4, 16)$ QAM, WGC receiver for 64-ary DPSK, and ML receiver for coherent 64-ary rectangular QAM with $\widehat{\delta}_{sd} = \widehat{\delta}_{sr} = \widehat{\delta}_{rd} = 0$, $m = 1.5$, $L = 2$, and $(\Upsilon_{sd}, \Upsilon_{sr}, \Upsilon_{rd}) = (1, 10, 1)$.

decision $\widehat{x}_n$ corresponding to

$$\max_{x_n \in \varkappa} \mathrm{Re}\{\vartheta x_n^*\} - \frac{|x_n|^2}{2}\left[\widehat{\gamma}_{sd}\widehat{z}_{sd}^2 + \sum_{l=1}^{L} \frac{\sigma_w^2 \alpha_l^2 \widehat{\gamma}_{sl}\widehat{\gamma}_{ld}\widehat{z}_{sl}^2\widehat{z}_{ld}^2}{1 + \sigma_w^2 \alpha_l^2 \widehat{\gamma}_{ld}\widehat{z}_{ld}^2}\right] \tag{16}$$

where the parameter ϑ is defined as

$$\vartheta \triangleq \frac{\sqrt{\widehat{\gamma}_{sd}}\widehat{h}_{sd}^* y_{sd,n}}{\sigma_w} + \sum_{l=1}^{L} \frac{\alpha_l \sqrt{\widehat{\gamma}_{sl}\widehat{\gamma}_{ld}}\widehat{h}_{sl}^* \widehat{h}_{ld}^* y_{ld,n}}{1 + \sigma_w^2 \alpha_l^2 \widehat{\gamma}_{ld}\widehat{z}_{ld}^2}$$

and $\varkappa \triangleq \{\pm\beta, \pm 3\beta, ..., \pm(I-1)\beta\}$ with $\beta = \sqrt{3/(2I^2-2)}$ denotes the I^2-ary rectangular QAM symbol set. Notably, the knowledge of the average SNR estimates $\widehat{\gamma}_{ij}$'s and the CIR estimates $\widehat{h}_{ij}$'s made at destination and the amplification factors α_l's used at relays are all required for realizing the coherent ML receiver. In addition to link SNR estimates, the requirement of link CIR estimates at destination complicates to a great extent the design of the coherent ML receiver when compared to both noncoherent EGC and WGC receivers.

With perfect link SNR estimation (i.e., $\widehat{\delta}_{sd} = \widehat{\delta}_{sl} = \widehat{\delta}_{ld} = 0$ for $l \in \mathcal{Z}_L^+$), the effect of incorrect CIR estimation is also shown in Fig. 4 to illustrate the sensitivity of the coherent receiver to CIR information. Specifically, we denote $\widehat{h}_{ij} \triangleq h_{ij} + \varepsilon_{ij}$ for $ij \in \{sd, \{sl\}_{l=1}^L, \{ld\}_{l=1}^L\}$ where the estimation error ε_{ij} is modeled as a CGRV with mean 0 and variance $\sigma_{\varepsilon,ij}^2$ [11] which is mutually independent and independent of the CIRs h_{ij}'s. For presentation simplicity, it is further assumed that all the CIR estimation errors have the same variance, i.e., $\sigma_{\varepsilon,ij} = \sigma_\varepsilon$. As shown in Fig. 4, the coherent ML receiver significantly outperforms noncoherent EGC and WGC receivers when the CIR estimation is correct. In the presence of CIR estimation errors, the coherent ML receiver degrades more significantly with larger errors and exhibits severe error-rate floors in high SNR region, where both noncoherent EGC and WGC receivers prevail.

VI. Conclusion

In this paper, the EGC receiver is developed to noncoherently combine received signals from direct and relay links and then demodulate DAPE QAM signals in the AF multiple relay system over independent Nakagami-m fading links. The EGC receiver for DAPE QAM is simpler to implement than noncoherent DPSK WGC and coherent QAM receivers since it is devoid of any CSI. Based on Beaulieu's series approach, an efficient BEP upper bound computation formula is analytically derived for the EGC receiver. The BEP upper bound is verified by simulation to be tight when the average link SNRs are sufficiently large. Performance results show that the EGC receiver for DAPE QAM performs much better than the WGC receiver for DPSK with the same constellation size. Moreover, the EGC receiver is shown to be less sensitive to SNR estimation errors than the WGC receiver for DPSK.

VII. Acknowledgement

This work was supported in part by the National Science Council of Taiwan under Contract NSC-100-2221-E-002-125-MY2 and by the National Taiwan University through the Excellent Research Project under Grant 10R80919-4.

References

[1] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inform. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[2] J. N. Laneman and G. W. Wornell, "Energy-efficient antenna sharing and relaying for wireless networks," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Chicago, IL, Sep. 23-28, 2000, vol. 1, pp. 7-12.

[3] T. Himsoon, W. P. Siriwongpairat, W. Su, and K. J. R. Liu, "Differential modulations for multinode cooperative communications," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2941–2956, Jul. 2008.

[4] S. S. Ikki and M. H. Ahmed, "Performance of cooperative diversity using equal gain combining (EGC) over Nakagami-m fading channels," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 557–562, Feb. 2009.

[5] W. Cho, R. Cao, and L. Yang, "Optimum resource allocation for amplify-and-forward relay networks with differential modulation," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5680–5691, Nov. 2008.

[6] C.-D. Chung, "Differentially amplitude and phase-encoded QAM for the correlated Rayleigh-fading channel with diversity reception," *IEEE Trans. Commun.*, vol. 45, no. 3, pp. 309–321, Mar. 1997.

[7] Y. Ma, Q. T. Zhang, R. Schober, and S. Pasupathy, "Diversity reception of DAPSK over generalized fading channels," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1834-1846, Jul. 2005.

[8] N. C. Beaulieu, "An infinite series for the computation of the complementary probability distribution function of a sum of independent random variables and its application to the sum of Rayleigh random variables," *IEEE Trans. Commun.*, vol. 38, no. 9, pp. 1463–1474, Sep. 1990.

[9] J. G. Proakis, *Digital Communications*, 4th ed. New York: McGraw-Hill, 2001.

[10] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, 9th ed. New York: Dover, 1970.

[11] S. Roy and P. Fortier, "Maximal-ratio combining architectures and performance with channel estimation based on a training sequence," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1154-1164, Jul. 2004.

[12] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed. San Diego, CA: Academic Press, 2000.

 ISBN: 978-1-61208-203-5

Full Rate Full Diversity Wireless Multicasting for Vehicle-to-Vehicle and Vehicle-to-Infrastructure Communications

Ali Ekşim
Center of Research for Advanced Technologies of Informatics and Information Security
TUBITAK-BILGEM, Gebze, Kocaeli, Turkey
E-mail: alieksim@uekae.tubitak.gov.tr

Mehmet E. Çelebi
Department of Electronic and Communication Engineering, Istanbul Technical University
Istanbul, Turkey
E-mail: mecelebi@itu.edu.tr

Abstract—**Multicasting is a spectrally efficient method for supporting group communication by allowing transmission of packets to multiple destinations using fewer resources. To incorporate cooperative diversity, Cooperative Extended Balanced Space-Time Block Codes (CEBSTBCs) have been proposed providing full diversity when one or more feedback bits are sent back via feedback channel. However, the CEBSTBCs are designed for cooperative unicast communication in the literature. This paper presents a novel wireless multicasting scheme which selects the optimum CEBSTBC for all vehicular users to support wireless multicast. The performance of the proposed scheme is investigated for not only vehicle-to-vehicle communication but also for vehicle-to-infrastructure cases. Extensive detailed simulations are performed to show the feasibility of full rate and full diversity multicast service provisioning in vehicular communications.**

Keywords-cooperative extended balanced space-time block coding; wireless multicasting; diversity; vehicular communications

I. INTRODUCTION

One of the space-time coding scheme is Orthogonal Space-Time Block Codes (OSTBCs), which provides full diversity advantage with low decoding complexity. The transmitted symbols are decoded separately using linear processing [1]. However, full diversity and full rate for more than two antennas cannot be achieved with OSTBCs. Several quasi-orthogonal STBCs that provide full rate at the expense of some loss in diversity [2],[3] and OSTBCs that provide full diversity with some loss in code rate [1], [4] have been proposed in the literature. In [5], full rate Balanced Space-Time Block Coding (BSTBC) have been proposed which achieve full diversity for arbitrary number of transmit antennas when one or more feedback bits are sent back via feedback channel. The main drawback of the BSTBC is limited coding gain. In [6-7], the Extended Balanced Space-Time Block Coding (EBSTBC) scheme has been proposed. In the EBSTBC, an arbitrary number of codes can be generated for improved coding gain.

Owing to insufficient antenna space, cost and hardware limitations, wireless users may not be able to support multiple transmit antennas. To overcome this difficulty, recently, researchers have been looking for methods to exploit spatial diversity using the antennas of different users in a network. This type of diversity is called the cooperative diversity [8] where virtual antenna arrays can be formed to overcome the drawback of channel correlation and space limitations of mobile unit. In addition, cooperative diversity reduces the required transmit power which leads to longer battery life and increases capacity in interference limited systems. The application of EBSTBCs into the cooperative communication is Cooperative Extended Balanced Space-Time Block Codes (CEBSTBCs), which was proposed in [7].

It is known that multicasting is an efficient method of supporting group communication as it allows for transmission of packets to multiple destinations using fewer network resources [9]. Along with the widespread deployment of wireless networks, the fast-improving capabilities of mobile devices, content and service providers are increasingly interested in supporting multicast communications over wireless networks.

Intelligent transportation systems (ITS) have recently attracted much attention from car manufacturers, road operators and standardization bodies. The primarily aim of ITS is to increase the road safety by means of vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication. Considerable effort has been dedicated to defining architectures, services and application scenarios for both V2V and V2I paradigms [10]. To the best our knowledge, there is no space-time block coding which achieves full rate and full diversity for more than one user for vehicular communication. In this paper, we propose a novel coding selection scheme for wireless multicasting. Extensive simulations are performed to show the feasibility of the full rate full diversity multicast service provisioning in V2V and V2I communication. In this regard, in the second section, the system models are described, in the third section, the CEBSTBCs are explained, in the fourth section, Multicast Cooperative Extended Balanced Space-Time Block Coding (MCEBSTBC) is presented, and in the last section, the results of the paper and the conclusion are given.

The following notation is used in the paper: The superscript * denotes the conjugate operation; Re{.} and Im{.} are the real and imaginary part of the argument, respectively. The operator $\lceil . \rceil$ rounds to the smallest integer greater or equal than its argument; the operator max(.) returns the largest of its operands and the min(.) returns the minimum of its operands.

 ISBN: 978-1-61208-203-5

II. SYSTEM MODELS

A. Vehicle-to-Vehicle System Model

The vehicle-to-vehicle system model consists of one source, N cooperative vehicles and L multicast vehicle users. All nodes are equipped with one single antenna. The Rayleigh channel model and the related second-order channel statistics originally proposed for a base station-to-mobile link fail to provide an accurate model for dynamic vehicle-to-vehicle link. Instead, the *cascaded (double)* Rayleigh fading channel model has been proposed [11-12], which provides a realistic description of an intervehicular channel where two Rayleigh fading processes are assumed to be generated by independent groups of scatterers around the two mobile terminals [13]. In the intervehicular system model, all channels are assumed to be frequency flat double Rayleigh fading channel. h_{sri} is the channel coefficient from the source vehicle to the ith cooperative vehicle (relay) and h_{ij} is the channel coefficient from the ith cooperative vehicle to the jth multicast vehicle user where i=1, 2,.., n and j=1, 2,.., L.

The channels are quasi-static, namely, the fading coefficients remain constant over the duration of one frame. Each multicast vehicle user is assumed to have perfect knowledge of its own channels. It is also assumed that the multicast users have no knowledge of the source vehicle-cooperative vehicle (relay) channels. Each cooperative vehicle is assumed to have perfect knowledge of its own source vehicle-cooperative vehicle channel. The cooperative vehicles employ amplify and forward protocol [8]. The noise is modeled as additive white Gaussian whose components are circular complex random variable with zero-mean and variance σ^2. P is the average transmitted power of the source vehicle and the cooperative vehicles. The source data bits are mapped by streams of y bits into M-ary phase shift keying (M-PSK) symbols, where $M=2^y$.

B. Vehicle-to-Infrastructure System Model

The vehicle-to-infrastructure system model is similar vehicle-to-vehicle system model except in the cooperative vehicle-multicast user channel part. In V2I system, cooperative vehicle to multicast user channels are assumed frequency flat Rayleigh fading channel where the channel gains are circularly complex Gaussian random variables and statistically independent from each other.

III. COOPERATIVE EXTENDED BALANCED SPACE-TIME BLOCK CODING

The Cooperative Extended Balanced Space-Time Block Coding (CEBSTBC) can be obtained when an extension matrix is multiplied with an Orthogonal Space-Time Block Coding [14-15]. Since Alamouti`s code is the only orthogonal code with rate one and minimum delay, the CEBSTBCs can be obtained as an extension of the Alamouti`s code [16].

$$\boldsymbol{C}=\boldsymbol{X}\boldsymbol{W}. \quad (1)$$

Here, $\boldsymbol{X}$ is the Alamouti's code and $\boldsymbol{W}$ is the 2xN matrix where $N\geq 2$ and the rank of $\boldsymbol{W}$ must be 2. The following example shows how to generate the CEBSTBCs for three transmitters. Consider the CEBSTBC pair with transmission matrix

$$\boldsymbol{C_1}=\begin{bmatrix} s_1 & s_2 & as_2 \\ -s_2^* & s_1^* & as_1^* \end{bmatrix} \quad (2)$$

where $a=e^{j2\pi m/q}$, q is the extension level and m=0, 1,…q-1. The columns and rows of $\boldsymbol{C_1}$ denote symbols transmitted from three cooperative relays in two signaling intervals, respectively. The matrix $\boldsymbol{C_1}$ is obtained from the Alamouti code using Equation (1) where

$$\boldsymbol{X}=\begin{bmatrix} s_1 & s_2 \\ -s_2^* & s_1^* \end{bmatrix} \quad \boldsymbol{W}=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & a \end{bmatrix}. \quad (3)$$

In this fashion, arbitrary number of the CEBSTBCs can be generated. It can be shown that the number of possible CEBSTBCs is $q^{N-2}(2^{N-1}-1)$ [7]. For that reason, the destination needs $N+d$ feedback bits ($N\geq 3$) to select any possible CEBSTBCs where $d=\lceil (N-2)\log_2 q \rceil - 1$. N-2 feedback bits are needed to achieve full diversity as in CBSTBCs [14]. The rest of the d+2 feedback bits provide an additional coding gain.

IV. MULTICAST COOPERATIVE EXTENDED BALANCED SPACE-TIME BLOCK CODING

Multicast Cooperative Extended Balanced Space-Time Block Coding (MCEBSTBC) can be obtained when an optimum CEBSTBC is selected for all multicast users. The MCEBSTBC contains two phases: Multicast frame initialization phase and multicast transmission phase. In the first phase, the multicast users transmit their channel state information (CSI) to the selected multicast user and the selected multicast user selects the optimum CEBSTBC for all multicast users. This phase is shown in Figure 1. In Figure 2, multicast transmission phase is shown. In this phase, the source transmits data to the cooperative relays and the cooperative relays transmit to the multicast users according to selected the MCEBSTBC.

A. MCEBSTBC for Three Cooperative Vehicles

When three cooperative vehicles are present at the environment then, $\boldsymbol{C_1}$, $\boldsymbol{C_2}$ and $\boldsymbol{C_3}$ are available MCEBSTBC matrices. These matrices are

$$\boldsymbol{C_1}=\begin{bmatrix} s_1 & s_2 & as_2 \\ -s_2^* & s_1^* & as_1^* \end{bmatrix} \boldsymbol{C_2}=\begin{bmatrix} s_1 & s_2 & as_1 \\ -s_2^* & s_1^* & -as_2^* \end{bmatrix}$$

$$\boldsymbol{C_3}=\begin{bmatrix} s_1 & as_1 & s_2 \\ -s_2^* & -as_2^* & s_1^* \end{bmatrix}. \quad (4)$$

 ISBN: 978-1-61208-203-5

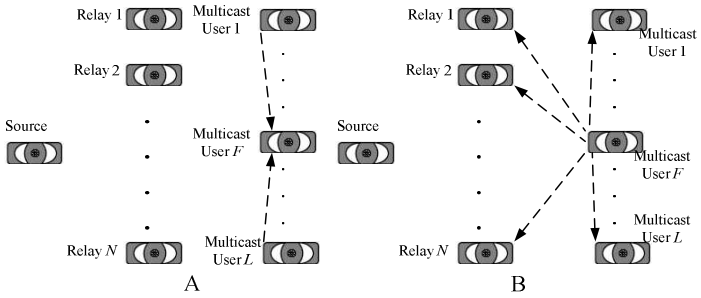


Figure 1. Multicast frame initialization phase of the MCEBSTBC: A) Channel coefficients are transmitted to the selected multicast vehicle user B) Selected code is transmitted both the cooperative vehicles and rest of the multicast vehicle users.

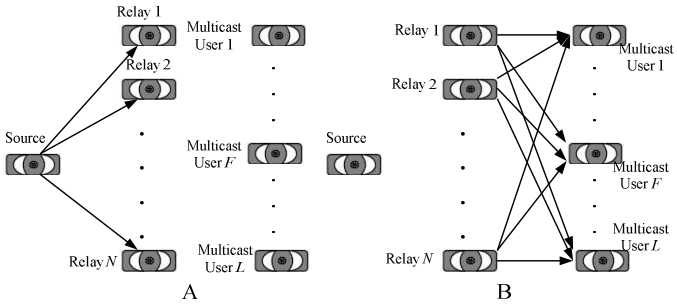


Figure 2. Multicast transmission phase of the MCEBSTBC: A) Broadcast phase B) Cooperation phase.

The selected multicast user picks the MCEBSTBC $\boldsymbol{C}_j$, j=1,2,3 that generates the optimum coding gain for all the multicast users. Two bits of feedback is needed to select the MCEBSTBC matrices and k bits of feedback is needed to select the feedback a where k =d+1. In [6-7], the optimum code is selected according to the single user channel coefficients. However, the optimum MCEBSTBC for all multicast users is selected according to the following maximin approach

$$A_1 = \max \begin{pmatrix} \min\left(\mathrm{Re}\left\{ah_{21}^* h_{31}\right\}, \mathrm{Re}\left\{ah_{22}^* h_{32}\right\}, ..., \mathrm{Re}\left\{ah_{2L}^* h_{3L}\right\}\right), \\ \min\left(\mathrm{Re}\left\{ah_{11}^* h_{31}\right\}, \mathrm{Re}\left\{ah_{12}^* h_{32}\right\}, ..., \mathrm{Re}\left\{ah_{1L}^* h_{3L}\right\}\right), \\ \min\left(\mathrm{Re}\left\{ah_{11}^* h_{21}\right\}, \mathrm{Re}\left\{ah_{12}^* h_{22}\right\}, ..., \mathrm{Re}\left\{ah_{1L}^* h_{2L}\right\}\right) \end{pmatrix}. \quad (5)$$

where a is selected to maximize the terms in the brackets [14]. The optimum MCEBSTBC is employed after combining, the observations at the jth vehicle multicast user, to obtain

$$\hat{s}_{i,j} = \sqrt{\frac{P}{3}}\left[\left|h_{1j}\right|^2 + \left|h_{2j}\right|^2 + \left|h_{3j}\right|^2 + 2A_1\right] s_i + \eta_{i,j}, \quad i = 1,2. \quad (6)$$

Here, $\hat{s}_{i,j}$ is the estimate of the ith symbol at the jth multicast vehicle user; $\eta_{1,j}$ and $\eta_{2,j}$ are the noise samples at the jth multicast vehicle user.

Fig. 3 shows the percentage of the channels that achieve full diversity for various multicast vehicle users when three cooperative vehicles are present in the environment. MCEBSTBC with one bit extension of feedback (MCEBSTBC (k=1)) achieves full diversity with only one user (unicast communication), since MCEBSTBC with one bit extension of feedback yields only 6 different codes. MCEBSTBC with two or more bit extension of feedback supports full diversity for two users. When five or more multicast users are present in the wireless environment, full diversity can be achieved in 70% or less of all possible channel conditions.

The following are the properties of the MEBSTBC for three cooperative relays:

i) One bit extension of feedback (k=1) cannot achieve full rate and full diversity for two multicast users.

ii) Two or more bit extension of feedback (k≥2) achieves full rate and full diversity for two multicast users.

iii) The full diversity can be achieved for an arbitrary number of multicast users, if the below inequality is satisfied for all possible channel conditions.

$$A_1 \geq 0. \quad (7)$$

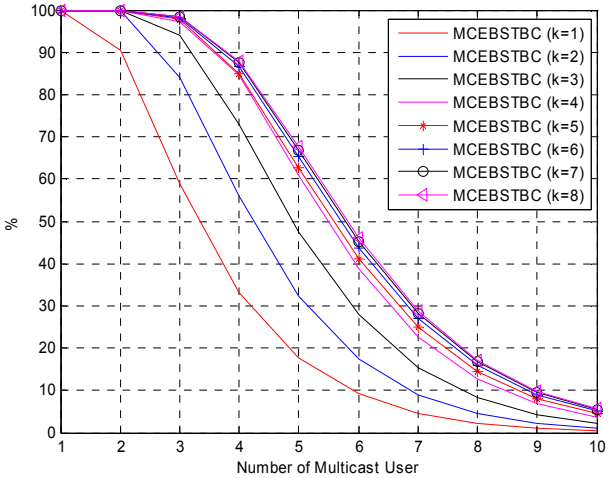


Figure 3. The percentage of the channels that achieve full diversity for various multicast vehicle users when three cooperative vehicles are present in the environment.

B. *MCEBSTBC for Four Cooperative Vehicles*

When four cooperative vehicles are present in the environment, available the MCEBSTBC matrices are

$$\boldsymbol{C}_1 = \begin{bmatrix} s_1 & as_1 & bs_1 & s_2 \\ -s_2^* & -as_2^* & -bs_2^* & s_1^* \end{bmatrix} \boldsymbol{C}_2 = \begin{bmatrix} s_1 & as_1 & s_2 & bs_1 \\ -s_2^* & -as_2^* & s_1^* & -bs_2^* \end{bmatrix}$$

$$\boldsymbol{C}_3 = \begin{bmatrix} s_1 & s_2 & as_1 & bs_1 \\ -s_2^* & s_1^* & -as_2^* & -bs_2^* \end{bmatrix} \boldsymbol{C}_4 = \begin{bmatrix} s_1 & s_2 & as_2 & bs_2 \\ -s_2^* & s_1^* & as_1^* & bs_1^* \end{bmatrix} \quad (8)$$

$$\boldsymbol{C}_5 = \begin{bmatrix} s_1 & as_1 & s_2 & bs_2 \\ -s_2^* & -as_2^* & s_1^* & bs_1^* \end{bmatrix} \boldsymbol{C}_6 = \begin{bmatrix} s_1 & s_2 & as_1 & bs_2 \\ -s_2^* & s_1^* & -as_2^* & bs_1^* \end{bmatrix}$$

$$\boldsymbol{C}_7 = \begin{bmatrix} s_1 & s_2 & as_2 & bs_1 \\ -s_2^* & s_1^* & as_1^* & -bs_2^* \end{bmatrix}.$$

$$A_2 = \max\left(\begin{array}{l}
\min\left(\left[\mathrm{Re}\left\{ah_{11}^{*}h_{21}\right\}+\mathrm{Re}\left\{bh_{11}^{*}h_{31}\right\}+\mathrm{Re}\left\{a^{*}bh_{21}^{*}h_{31}\right\}\right],\ldots,\left[\mathrm{Re}\left\{ah_{1L}^{*}h_{2L}\right\}+\mathrm{Re}\left\{bh_{1L}^{*}h_{3L}\right\}+\mathrm{Re}\left\{a^{*}bh_{2L}^{*}h_{3L}\right\}\right]\right),\\
\min\left(\left[\mathrm{Re}\left\{ah_{11}^{*}h_{21}\right\}+\mathrm{Re}\left\{bh_{11}^{*}h_{41}\right\}+\mathrm{Re}\left\{a^{*}bh_{21}^{*}h_{41}\right\}\right],\ldots,\left[\mathrm{Re}\left\{ah_{1L}^{*}h_{2L}\right\}+\mathrm{Re}\left\{bh_{1L}^{*}h_{4L}\right\}+\mathrm{Re}\left\{a^{*}bh_{2L}^{*}h_{4L}\right\}\right]\right),\\
\min\left(\left[\mathrm{Re}\left\{ah_{11}^{*}h_{31}\right\}+\mathrm{Re}\left\{bh_{11}^{*}h_{41}\right\}+\mathrm{Re}\left\{a^{*}bh_{31}^{*}h_{41}\right\}\right],\ldots,\left[\mathrm{Re}\left\{ah_{1L}^{*}h_{3L}\right\}+\mathrm{Re}\left\{bh_{1L}^{*}h_{4L}\right\}+\mathrm{Re}\left\{a^{*}bh_{3L}^{*}h_{4L}\right\}\right]\right),\\
\min\left(\left[\mathrm{Re}\left\{ah_{21}^{*}h_{31}\right\}+\mathrm{Re}\left\{bh_{21}^{*}h_{41}\right\}+\mathrm{Re}\left\{a^{*}bh_{31}^{*}h_{41}\right\}\right],\ldots,\left[\mathrm{Re}\left\{ah_{2L}^{*}h_{3L}\right\}+\mathrm{Re}\left\{bh_{2L}^{*}h_{4L}\right\}+\mathrm{Re}\left\{a^{*}bh_{3L}^{*}h_{4L}\right\}\right]\right),\\
\min\left(\left[\mathrm{Re}\left\{ah_{11}^{*}h_{21}\right\}+\mathrm{Re}\left\{bh_{31}^{*}h_{41}\right\}\right],\ldots,\left[\mathrm{Re}\left\{ah_{1L}^{*}h_{2L}\right\}+\mathrm{Re}\left\{bh_{3L}^{*}h_{4L}\right\}\right]\right),\\
\min\left(\left[\mathrm{Re}\left\{ah_{11}^{*}h_{31}\right\}+\mathrm{Re}\left\{bh_{21}^{*}h_{41}\right\}\right],\ldots,\left[\mathrm{Re}\left\{ah_{1L}^{*}h_{3L}\right\}+\mathrm{Re}\left\{bh_{2L}^{*}h_{4L}\right\}\right]\right),\\
\min\left(\left[\mathrm{Re}\left\{ah_{21}^{*}h_{31}\right\}+\mathrm{Re}\left\{bh_{11}^{*}h_{41}\right\}\right],\ldots,\left[\mathrm{Re}\left\{ah_{2L}^{*}h_{3L}\right\}+\mathrm{Re}\left\{bh_{1L}^{*}h_{4L}\right\}\right]\right)
\end{array}\right). \quad (9)$$

The optimum MCEBSTBC for all multicast vehicle users is chosen according to Equation (9) where a and b are selected to maximize the terms in the brackets [14]. After combining the observations the estimates are obtained as shown in Equation (10). Here $\eta_{1,j}$ and $\eta_{2,j}$ are the noise samples at the jth mobile user.

$$\hat{s}_{i,j} = \frac{\sqrt{P}}{2}\left[\left|h_1\right|^2+\left|h_2\right|^2+\left|h_3\right|^2+\left|h_4\right|^2+2A_2\right]s_i+\eta_{i,j} \quad (10)$$

where i=1,2.

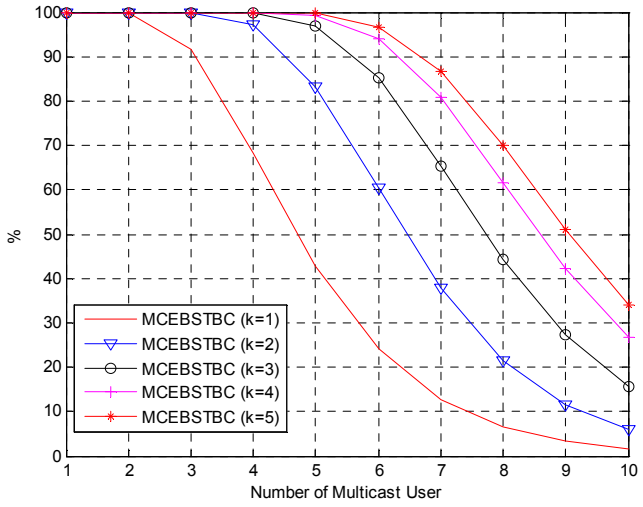


Figure 4. Percentage of all possible channel conditions that achieves full diversity for various multicast vehicle users when four cooperative vehicles are present in the environment.

Fig. 4 shows the percentage of channels that achieve full diversity for various multicast users when four cooperative vehicles are present in the wireless environment. MCEBSTBC with one bit extension of feedback (MCEBSTBC (k=1)) achieves full diversity and full rate for two multicast users. When eight or more multicast users are present in the wireless environment and up to five bit extension of feedback is available, full diversity can be achieved 70% or less of all possible channel conditions.

The following are the properties of the MEBSTBC for four or more cooperative relays:

i) One bit extension of feedback (k=1) can achieve full rate and full diversity for two multicast users.

ii) When four cooperative relays are present in the wireless environment, full diversity can be achieved for an arbitrary number of multicast users, if the inequality of Equation (11) is satisfied for all possible channel conditions.

$$A_2 \geq 0. \quad (11)$$

V. PERFORMANCE EVALUATIONS

The bit error probabilities of the MCEBSTBC are evaluated for quaternary phase-shift keying (QPSK) modulation by computer simulations. The frame length is 128 symbol duration. The source vehicle-cooperative vehicle (relay) channels are better quality in signal-to-noise ratio (SNR) than cooperative vehicle-multicast vehicle user channels whose difference is quantified by differential signal-to-noise ratio (DSNR). In the Figures 5-8, DSNR is assumed to be 25 dB for three and four cooperative vehicles. For comparison, the bit error rate (BER) curve of the unicast CEBSTBC [7] is also included in Figure 5-8.

Figure 5 presents the bit error probabilities of the MCEBSTBC with four bits extension of feedback for three cooperative vehicles and various numbers of multicast vehicle users. It can be seen from the Figure 5 that the full diversity cannot be achieved for more than four multicast users since the slope of the curves is decreased. Compared to the MCEBSTBC with 2 multicast users (2 Mult. MCEBSTBC (k=4)), the CEBSTBC with four bits extension of feedback (Unicast CEBSTBC (k=4) [7]) has a SNR advantage of only 0.54 dB for a BER value of $1\text{x}10^{-3}$. However, the MCEBSTBC with 2 multicast users (2 Mult. MCEBSTBC (k=4)) provides better performance than the CEBSTBC with one bit extension of feedback (Unicast CEBSTBC (k=1) [7]) and the system transmission rate is doubled. Relative to the MCEBSTBC with 3 multicast users (3 Mult. MCEBSTBC (k=4)), the MCEBSTBC with 4 multicast users (4 Mult. MCEBSTBC (k=4)), and the MCEBSTBC with 5 multicast users (5 Mult. MCEBSTBC

 ISBN: 978-1-61208-203-5

(k=4)), the CEBSTBCs with four bit extension of feedback (Unicast CEBSTBC (k=4) [7]) has a SNR advantage of merely 1.1 dB, 1.7 dB, and 2.45 dB, respectively. The proposed MCEBSTBC sacrifices some coding gain to utilize system resources efficiently.

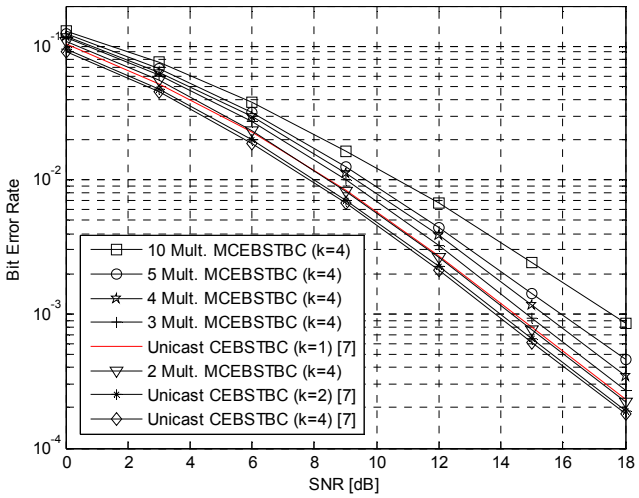


Fig. 5. BER of the CEBSTBC and the MCEBSTBC when three cooperative vehicles are present.

Figure 6 presents the bit error probabilities of the MCEBSTBC with four bits extension of feedback for four cooperative relays and various numbers of multicast users. It can be seen from the Figure 7 that the full diversity can be achieved for five multicast users since the slope of the curves does not change. Compared to the CEBSTBC with four bits extension of feedback (Unicast CEBSTBC (k=4) [7]), the MCEBSTBC with 2 multicast users (2 Mult. MCEBSTBC (k=4)) has a SNR advantage of just 0.77 dB for a BER value of $1x10^{-4}$. However, the MCEBSTBC with 3 multicast users (3 Mult. MCEBSTBC (k=4)) provides just 0.25 dB worse performance than the CEBSTBC with one bit extension of feedback (Unicast CEBSTBC (k=1) [7]) and the system transmission rate is tripled. In comparison the MCEBSTBC with 3 multicast users (3 Mult. MCEBSTBC (k=4)), the MCEBSTBC with 4 multicast users (4 Mult. MCEBSTBC (k=4)), and the MCEBSTBC with 5 multicast users (5 Mult. MCEBSTBC (k=4)), the CEBSTBCs with four bit extension of feedback (Unicast CEBSTBC (k=4) [7]) has a SNR advantage of only 1.37 dB, 1.88 dB, and 2.21 dB, respectively. Once again, the proposed MCEBSTBC sacrifices a slight coding gain but the system transmission rate is increased L times.

In the sequel, we simulate the cooperative V2I communication. In this scenario, the multicast users are at the infrastructure and the cooperative vehicle-multicast users´ channels are Rayleigh fading. Figure 7 depicts the bit error probabilities of the MCEBSTBC with four bits extension of feedback for three cooperative vehicles and various numbers of multicast users. It can be seen from the Figure 7 that the full diversity cannot be achieved more than four multicast users since the slope of curves is decreased. Table 1 presents required SNR values for a BER value of $1x10^{-3}$. It can be easily seen that the proposed MCEBSTBC sacrifices some coding gain to utilize system resources efficiently.

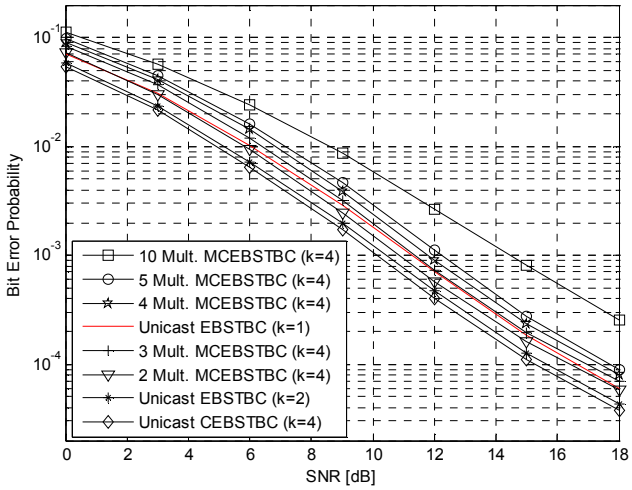


Fig. 6. BER of the CEBSTBC and the MCEBSTBC when four cooperative vehicles are present.

TABLE I. REQUIRED SNR VALUES FOR BER VALUE OF $1x10^{-3}$

Unicast/Multicast Transmission Schemes	Required SNR Values
10 Mult. MCEBSTBC (k=4)	13.40 dB
5 Mult. MCEBSTBC (k=4)	11.76 dB
4 Mult. MCEBSTBC (k=4)	11.28 dB
3 Mult. MCEBSTBC (k=4)	10.79 dB
Unicast CEBSTBC (k=1) [7]	10.35 dB
2 Mult. MCEBSTBC (k=4)	10.33 dB
Unicast CEBSTBC (k=2) [7]	9.92 dB
Unicast CEBSTBC (k=4) [7]	9.78 dB

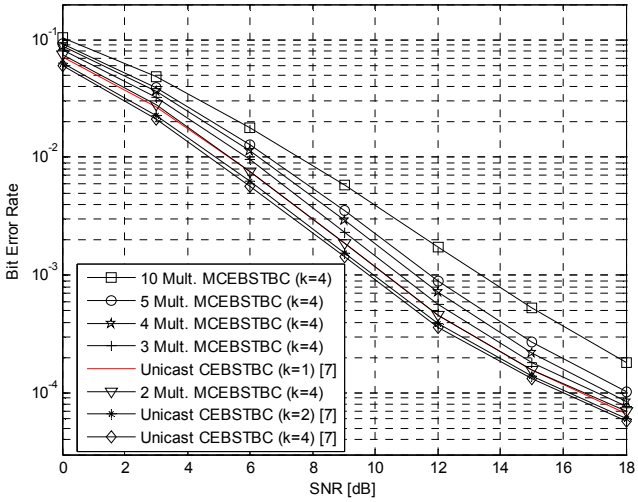


Fig. 7. BER of the CEBSTBC and the MCEBSTBC when three cooperative vehicles are present.

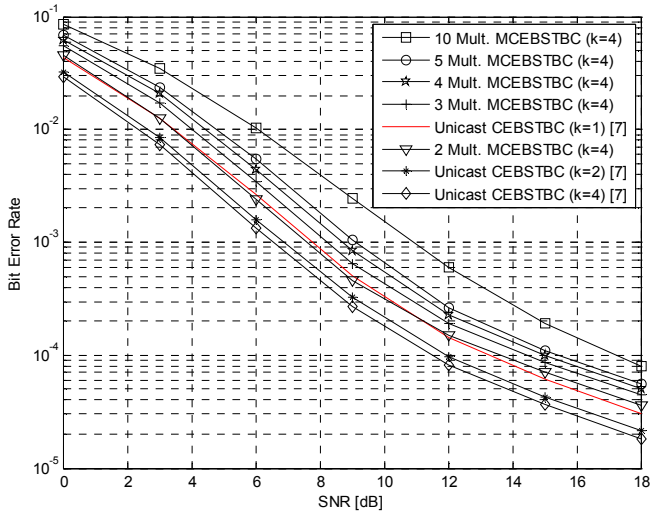


Fig. 8. BER of the CEBSTBC and the MCEBSTBC when four cooperative vehicles are present.

Figure 8 presents the bit error probabilities of the MCEBSTBC with four bits extension of feedback for four cooperative vehicles and various numbers of multicast users. It can be seen from the Figure 8 that the full diversity can be achieved for five multicast users since the slope of the curves does not change. Table 2 presents required SNR values for a BER value of $1x10^{-3}$. Once again, the proposed MCEBSTBC sacrifices a slight coding gain but the system transmission rate is increased L times.

TABLE II. REQUIRED SNR VALUES FOR BER VALUE OF $1x10^{-3}$

Unicast/Multicast Transmission Schemes	Required SNR Values
10 Mult. MCEBSTBC (k=4)	10.93 dB
5 Mult. MCEBSTBC (k=4)	9.10 dB
4 Mult. MCEBSTBC (k=4)	8.68 dB
3 Mult. MCEBSTBC (k=4)	8.22 dB
Unicast CEBSTBC (k=1) [7]	7.77 dB
2 Mult. MCEBSTBC (k=4)	7.58 dB
Unicast CEBSTBC (k=2) [7]	6.88 dB
Unicast CEBSTBC (k=4) [7]	6.55 dB

VI. CONCLUSION

In this paper, full rate and full diversity multicast service provisioning in V2V and V2I communications was analyzed and simulated. It has been observed that compared to the unicast CEBSTBC, the MCEBSTBC does not utilize all available codes and employs optimum CEBSTBC for all multicast vehicle users. This optimization sacrifices a slight coding gain to utilize system resources efficiently. Namely, by using the MCEBSTBC, the system transmission rate is increased in proportion to the number of multicast users. The larger cooperative vehicles present at the wireless environment, the fuller diversity full rate wireless multicasting can be achieved. The proposed multicast technique might be implemented easily in IEEE 802.11p [17] which defines enhancements to 802.11 required to support ITS applications [17].

REFERENCES

[1] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," IEEE Trans. on Information Theory, vol. 45, pp. 1456-1467, 1999.

[2] H. Jafarkhani, "A quasi-orthogonal space-time block code," IEEE Trans. Commun., vol. 49, pp. 1-4, 2001.

[3] O. Tirkkonen, and A. Hottinen, "Complex space-time block codes for four Tx antennas," Proc. IEEE GLOBECOM, IEEE Press, December 2000, pp. 1005-1009.

[4] W. Su, and X. G. Xia, "On space-time block codes from complex orthogonal designs," Wirel. Pers. Commun., vol. 25, pp. 1-26, 2003.

[5] M. E. Çelebi, S. Şahin, and Ü. Aygölü, "Increasing diversity with feedback: Balanced space-time block coding," Proc. IEEE ICC, IEEE Press, June 2006, pp. 4836-4841.

[6] A. Ekşim, and M. E. Çelebi, "Extended cooperative balanced space-time block coding for increased efficiency in wireless sensor networks (Work in Progress)," Networking 2009, vol. 5550, pp. 456–467, May 2009.

[7] A. Ekşim, and M. E. Çelebi, "Extended balanced space-time block coding for wireless communications," IET Signal Processing, vol. 3, pp. 476-484, November 2009.

[8] J. N. Laneman, G. W. Wornell, and D. N. C. Tse, "An efficient protocol for realizing cooperative diversity in wireless networks," Proc. IEEE ISIT, June 2001, pp. 294.

[9] U. Varshney, "Multicast over wireless networks," Wirel. Commun. Mob. Comput, vol. 2, pp. 667–692, 2002.

[10] P. Belanovic, D. Valerio, A. Paier, T. Zemen, F.Ricciato, and C. F. Mecklenbrauker, "On Wireless Links for Vehicle-to-Infrastructure Communications," IEEE Transactions on Vehicular Technology, vol. 59, pp. 269-282, 2010.

[11] I. Z. Kovacs, "Radio channel characterisation for private mobile radio systems: Mobile-to-mobile radio link investigations," Ph.D. dissertation, Aalborg Univ., Aalborg, Denmark, Sep. 2002.

[12] G. K. Karagiannidis, T. A. Tsiftsis, and N. C. Sagias, "A closedform upper-bound for the distribution of the weighted sum of Rayleigh variates," IEEE Commun. Lett., vol. 9, no. 7, pp. 589–591, Jul. 2005.

[13] H. Ilhan, M. Uysal, Ibrahim Altunbaş, "Cooperative Diversity for Intervehicular Communication: Performance Analysis and Optimization", IEEE Transactions on Vehicular Technology, vol. 58, No. 7, September 2009.

[14] A. Ekşim, "Extended Balanced Space-Time Block Coding in Wireless Networks," PhD. Thesis, Istanbul Technical University, 2011.

[15] A. Ekşim, and M. E. Çelebi, "Performance Improvement of Binary Sensor Based Statistical STBC Cooperative Diversity Using Limited Feedback," IETE Technical Review, vol. 27, pp. 60-67, Jan-Feb 2010.

[16] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," IEEE J. Select. Areas Commun., 16, pp. 1451-1458, 1998.

[17] IEEE 802.11p standard, 15 July 2010. Available at: http://standards.ieee.org/getieee802/download/802.11p-2010.pdf [retrieved: April, 2012].

 ISBN: 978-1-61208-203-5

Analysis of Statistical Time-access Fairness Index of Opportunistic Feedback Fair Scheduler

Fumio Ishizaki
Department of Systems Design and Engineering
Nanzan University
27 Seirei, Seto, Aichi 489-0863, Japan
Email: fumio@ieee.org

***Abstract*—Since the utilization of multiuser diversity in wireless networks can increase the information theoretic capacity of the overall system, much attention has been paid to schedulers exploiting multiuser diversity. However, packet schedulers exploiting multiuser diversity have a disadvantage of consuming the bandwidth for the feedback load. From a view of feedback reduction, the opportunistic feedback fair scheduler is considered as an attractive choice among schedulers exploiting multiuser diversity. In this paper, considering the statistical time-access fairness index (STAFI) as a measure of short term fairness, we study the short term fairness provided by the opportunistic feedback fair scheduler. Numerical results display that the threshold value of the scheduler greatly affects the properties of its short term fairness.**

***Keywords*-Opportunistic feedback fair scheduler; Short term fairness; Statistical time-access fairness index**

I. INTRODUCTION

Multiuser diversity [1] is a diversity existing between the channel states of different users in wireless networks. Since packet schedulers exploiting multiuser diversity have an advantage of increasing the information theoretic capacity of the overall system, much attention has been paid to such schedulers (see, e.g., [2], [3], [4], [5], [6], [7], [8] and references therein). However, packet schedulers exploiting multiuser diversity also have a disadvantage of consuming the bandwidth for the feedback load, defined as the amount of channel information that needs to be fed back from MSs (mobile stations) to BS (base station). In addition, it is known that there exists a tradeoff between the information theoretic capacity and fairness achieved by schedulers exploiting multiuser diversity [9]. Therefore, when we consider a scheduler exploiting multiuser diversity, we should take its feedback load and fairness as well as performance gain into account.

To reduce the feedback load while still having the performance gain, several schedulers have been proposed and studied. The one-bit feedback fair scheduler [10], [11], [12], [13] is an example of such schedulers. Under the one-bit feedback fair scheduling, the *normalized* received SNR (Signal-to-Noise Ratio) values of MSs (instead of the received SNR values) are considered. Each MS feeds back one-bit information to BS, only when its *normalized* received SNR is greater than or equal to a predetermined threshold. By doing so, the one-bit feedback fair scheduler can reduce the feedback load from MSs to BS and achieve the ideal *long term fairness*, while having considerable performance gain. However, the one-bit feedback fair scheduler still has a difficulty for the feedback load. The difficulty is that the feedback load of the one-bit feedback fair scheduler linearly increases with the number of MSs, although the performance gain for the capacity also grows as the number of MSs becomes large [14]. This may degrade the scalability of the one-bit feedback fair scheduler. One way to overcome the difficulty against the scalability is to introduce a random access-based feedback scheme. As a scheduler with random access-based feedback scheme, Tang and Heath [7] proposed the opportunistic feedback scheduler. Under the opportunistic feedback scheduling, the feedback resources are random access minislots. MSs transmit feedback information with some probability in each minislot only when their SNR values are greater than or equal to a predetermined threshold. Contrary to the one-bit feedback fair scheduler, the feedback load of the opportunistic feedback scheduler is independent of the number of MSs.

The fairness of scheduler is classified into short term fairness and long term fairness [15], [16]. While long term fairness governs the long run performances such as long run average throughput of individual MSs, short term fairness greatly affects the packet level performances such as delay and loss probability of individual MSs. Since the packet level performances of individual MSs are basic measures of QoS, it is important to examine the short term fairness of scheduler in terms of QoS guarantees.

As a measure of short term fairness, the proportional fairness index is usually considered in wireline networks. The proportional fairness index characterizes the service discrepancy *in bits* between two flows over any time interval during which the two flows are continuously backlogged. However, for the following two reasons, the proportional fairness index is not suitable for wireless networks. First, the proportional fairness index considers the hard deterministic guarantee, and it does not take randomness inherent in the wireless channel conditions into account. Second, the proportional fairness index considers fairness of users' throughputs rather than channel access times, although users can transmit at different rates depending on their current channel quality in wireless

This research was supported by Nanzan University Pache Research Subsidy I-A-2 for the 2012 academic year.

 ISBN: 978-1-61208-203-5

networks. Liu et al. [17] then consider modifications to the proportional fairness index for short term fairness index in wireless networks. By considering the service in *time* (instead of the service in bits) and a statistical fairness guarantee (instead of the hard deterministic fairness guarantee), they propose a *statistical time-access fairness index* (STAFI) defined as

$$\mathrm{P}\left(\left|\frac{\alpha^{(i)}(t_1,t_2)}{\phi^{(i)}}-\frac{\alpha^{(j)}(t_1,t_2)}{\phi^{(j)}}\right|\geq x\right)\leq f^{(i,j)}(x), \qquad (1)$$

where $\alpha^{(i)}(t_1,t_2)$ denotes the service *in time* that flow i receives during $[t_1,t_2)$, ϕ_i denotes the assigned weight for flow i and $f^{(i,j)}(x)$ is a probability distribution which may depend on i and j.

In this paper, we focus on the short term fairness of the opportunistic feedback fair scheduler. We study the STAFI of the scheduler to investigate its short term fairness properties. In particular, we consider the STAFI where the assigned weights ϕ_i in (1) are all equal to one. Since the *normalized* SNR processes of MSs are considered and the normalized SNRs of MSs are i.i.d. (independent and identically distributed) under the opportunistic feedback fair scheduling, the opportunistic feedback fair scheduler provides an ideal long term fairness property [5]. However, as far as the author's best knowledge, there is no study on the short term fairness properties of the opportunistic feedback fair scheduler, although the packet level performances of individual MSs are strongly affected by the short term fairness.

The remainder of this paper is organized as follows. In Section II, we describe a system model considered in this paper. We assume that the wireless channel process for each user is modeled by a discrete-time two-state Markov chain. We analyze the STAFI of the opportunistic feedback fair scheduler in Section III. We also develop a numerical method to calculate the exact value of the STAFI by using the inverse discrete FFT method [18]. Section IV provides numerical results to investigate the properties of the short term fairness provided by the opportunistic feedback fair scheduler. Conclusion is drawn in Section V.

II. System model

In this paper, we consider a wireless network consisting of a BS and K MSs. We suppose that the BS employs the opportunistic feedback fair scheduler for downlink transmission from the BS to the MSs [7]. In this paper, considering the STAFI as a measure of short term fairness, we study the properties of short term fairness provided by the opportunistic feedback fair scheduler for the downlink transmission.

We assume that the downlink channel of MS i $(i = 1,\ldots,K)$ is described by a Rayleigh fading channel model. Time axis is divided into frames of equal size T_f (sec) and time index is given by $t = 0,1,2,\cdots$. The frame duration T_f is considered to be the unit time in our model. Then, the received SNR process $\{z^{(i)}(t)\}$ $(t = 0,1,\ldots)$ of MS i $(i = 1,\ldots,K)$ is described as a discrete-time stochastic process. We assume that the received SNR processes of the K MSs are independent with each other.

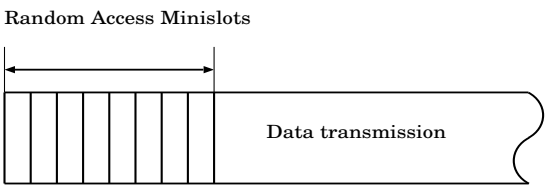


Fig. 1. Uplink frame structure

Without loss of generality, we consider that MS 1 and MS 2 are tagged users and all the other MSs are background users for the STAFI. More specifically, we assume that for $i = 3,\ldots,K$, the received SNR process $\{z^{(i)}(t)\}$ is a stationary process. But we do not assume the stationarity of $\{z^{(i)}(t)\}$ for $i = 1,2$. When the received SNR process $\{z^{(i)}(t)\}$ is stationary, $z^{(i)}(t)$ at time t is according to the following exponential distribution:

$$\mathrm{P}\{z^{(i)}(t)\leq x\} = 1-\exp(-x/\bar{z}^{(i)}), \qquad (2)$$

where $\bar{z}^{(i)}$ denotes the average received SNR of MS i and is defined by $\bar{z}^{(i)} = \mathrm{E}[z^{(i)}(t)]$.

A. Opportunistic feedback fair scheduler

Under the opportunistic feedback fair scheduling, the *normalized* SNR processes of MSs are considered, where the normalized SNR process is defined by the process $\{z^{(i)}(t)/\bar{z}^{(i)}\}$ $(i = 1,\ldots,K)$. To reduce the feedback load, each MS quantizes or partitions the entire normalized SNR range into two grades with threshold denoted by γ_1. We assume that the threshold γ_1 is *a priori* determined. If $z^{(i)}(t)/\bar{z}^{(i)} < \gamma_1$, we say that the wireless channel state of MS i is in state 0 at time t. If $z^{(i)}(t)/\bar{z}^{(i)} \geq \gamma_1$, we say that the wireless channel state of MS i is in state 1 at time t. We assume that perfect channel estimation is possible at each MS and each MS knows its average SNR $\bar{z}^{(i)}$ $(i = 1,\ldots,K)$. Then MS i can determine the grade of its channel to the BS with the knowledge of its normalized SNR.

We suppose that the opportunistic feedback fair scheduler is employed in a frequency-division-duplex (FDD) system. In the FDD system, at the beginning of the downlink frame, the BS broadcasts a message containing the information for opportunistic feedback to all the MSs. N minislots in an uplink frame for random access feedback follow the downlink message as illustrated in Figure 1. We assume that the number of minislots N is fixed.

The opportunistic feedback fair scheduler then operates as follows:

- At every time t, MS i estimates its received normalized SNR $z^{(i)}(t)/\bar{z}^{(i)}$ and examines if $z^{(i)}(t)/\bar{z}^{(i)}$ is greater than or equal to the threshold γ_1.
- If the normalized SNR of MS i $z^{(i)}(t)/\bar{z}^{(i)}$ $(i = 1,\ldots,K)$ is greater than or equal to the threshold γ_1 (i.e., if the wireless channel state of MS i is in state 1), MS i attempts to transmit feedback information to the

 ISBN: 978-1-61208-203-5

MS with a probability u in every minislot. We hereafter call the probability u the feedback probability.
- Otherwise (i.e., if the wireless channel state of MS i is in state 0), MS i does not feedback any information to the BS in the random access minislots.
- The feedback information can be fed back to the BS if and only if one MS attempts to transmit feedback information in the minislot. Otherwise, either a collision happens or there is no MS to feed back.
- If multiple MSs successfully feedback during the random access period consisting of N minislots, the BS randomly selects one of the successful MSs.
- If there is no successful feedback in all N minislots, the BS randomly selects one MS among all the K MSs.
- The scheduling is performed frame-by-frame.

We assume that the random access attempts are independent among MSs and also independent among random access minislots.

B. Wireless channel model

In this subsection, we consider a wireless channel state process of MS i $(i = 1, \ldots, K)$. Let $\{s^{(i)}(t)\}$ $(t = 0, 1, \ldots; i = 1, \ldots, K)$ denote the wireless channel state process of MS i, where $s^{(i)}(t) = 1$ if $z^{(i)}(t)/\bar{z}^{(i)} \geq \gamma_1$ and $s^{(i)}(t) = 0$ otherwise. We assume that the channel state process $\{s^{(i)}(t)\}$ $(t = 0, 1, \ldots; i = 1, \ldots, K)$ of MS i is well described by a discrete-time 2-state Markov chain [15], [19]. We further assume that for $i = 3, \ldots, K$, the Markov chain $\{s^{(i)}(t)\}$ is stationary from the assumption of the stationarity of the received SNR process $\{z^{(i)}(t)\}$ for $i = 3, \ldots, K$. On the other hand, for $i = 1, 2$, we do not assume the stationarity of $\{s^{(i)}(t)\}$.

Let $\boldsymbol{P} = (p_{i,j})$ $(i, j = 0, 1)$ denote the transition probability matrix of the 2-state Markov chain. The transition probability matrix $\boldsymbol{P}$ is determined as follows (for the detailed derivation of the transition probabilities, see [19]). We first consider the level crossing rate $\chi(\gamma)$ of the received normalized SNR at γ given by [20]

$$\chi(\gamma) = \sqrt{2\pi\gamma} f_d \exp(-\gamma), \tag{3}$$

where f_d denotes the mobility-induced Doppler spread of MSs and we assume that for all the MSs, the mobility-induced Doppler spreads are identical.

For MS i $(i = 3, \ldots, K)$, we next consider the stationary probability vector $\boldsymbol{s} = (s_0, s_1)$ of the 2-state discrete-time Markov chain $\{s^{(i)}(t)\}$. Note here that for the MSs $(i = 3, \ldots, K)$, the channel state processes have the same stationary probability vector due to the normalization of the received SNRs. From (2), the stationary probability vector is given by

$$s_0 = 1 - e^{-\gamma_1}, \quad s_1 = e^{-\gamma_1}. \tag{4}$$

The state transition probabilities are then determined by

$$p_{0,1} = \frac{\chi(\gamma_1) T_f}{s_0}, \quad p_{1,0} = \frac{\chi(\gamma_1) T_f}{s_1}, \tag{5}$$

$$p_{0,0} = 1 - p_{0,1}, \quad p_{1,1} = 1 - p_{1,0}, \tag{6}$$

where s_i $(i = 0, 1)$ and $\chi(\gamma_1)$ are given by (4) and (3), respectively. (5) and (6) determine the transition probability matrix $\boldsymbol{P}$ of the 2-state Markov chain, whose stationary probability vector is given by (4).

III. Analysis

In this section, we analyze the STAFI between MS 1 and and MS 2, which are tagged users.

Let $c^{(i)}(t)$ $(i = 1, \ldots, K; t = 0, 1, \ldots)$ denote a random variable representing the amount of service of MS i at time t, i.e., $c^{(i)}(t) = 1$ when the opportunistic feedback fair scheduler selects MS i for downlink transmission at time t, and $c^{(i)}(t) = 0$ otherwise. The amount service $\alpha^{(i)}(t_0 t_0 + n)$ for MS i in $[t_0, t_0 + n)$ is then expressed as

$$\alpha^{(i)}(t_0, t_0 + n) = \sum_{t=t_0}^{t_0+n-1} c^{(i)}(t).$$

In this paper, we hereafter consider only the cases where $t_0 = 0$, because we focus on the transient properties of the short term fairness of the scheduler. Let $\beta^{(i,j)}(n)$ $(i, j = 1, \ldots, K; n = 0, 1, \ldots)$ denote the difference between the amount service for MS i and that for MS j in $[t_0, t_0 + n)$. $\beta^{(i,j)}(n)$ is given by

$$\beta^{(i,j)}(n) = |\alpha^{(i)}(0, n) - \alpha^{(j)}(0, n)|.$$

We are now ready to provide an expression of the STAFI of the scheduler. Let $G_n(x)$ $(n = 1, 2, \ldots)$ denote the STAFI. $G_n(x)$ is defined by

$$\begin{aligned} G_n(x) &= \mathrm{P}(\beta^{(1,2)}(n) \geq x) \\ &= \mathrm{P}(|\alpha^{(1)}(0, n) - \alpha^{(2)}(0, n)| \geq x). \end{aligned}$$

We further define the probability mass function $g_n(x)$ $(n = 1, 2, \ldots)$ by

$$g_n(x) = \mathrm{P}(\beta^{(1,2)}(n) = x) = \mathrm{P}(|\alpha^{(1)}(0, n) - \alpha^{(2)}(0, n)| = x).$$

In what follows, we analyze the STAFI $G_n(x)$. For this purpose, we define some matrices and vectors. We first define a $(K-1) \times (K-1)$ matrix $\boldsymbol{R}$ by

$$\begin{aligned} [\boldsymbol{R}]_{i,j} &= \sum_{k=\max(0,i+j-K+2)}^{\min(i,j)} \binom{i}{k} p_{1,1}^k p_{1,0}^{i-k} \\ &\quad \cdot \binom{K-2-i}{j-k} p_{0,1}^{j-k} p_{0,0}^{K-2-i-j+k}, \end{aligned} \tag{7}$$

where $[\boldsymbol{R}]_{i,j}$ $(i, j = 0, \ldots, K-2)$ denotes the (i, j)th element of $\boldsymbol{R}$. Note that $\boldsymbol{R}$ is a transition probability matrix of the Markov chain $\{r(t)\}$ $(t = 0, 1, \ldots)$, where $r(t)$ is defined by $r(t) = \sum_{k=3}^{K} I(s^{(k)}(t) = 1)$. Thus, $[\boldsymbol{R}]_{i,j}$ denotes the conditional probability that j MSs among the $(K-2)$ MSs excluding MS 1 and MS 2 are in state 1 at time t given that i MSs among the $(K-2)$ MSs was in state 1 at time $t-1$. Let $\boldsymbol{r}$ denote the stationary probability vector of $\boldsymbol{R}$. The stationary probability vector $\boldsymbol{r}$ is given by

$$[\boldsymbol{r}]_j = \binom{K-2}{j} s_0^{K-2-j} s_1^j, \tag{8}$$

 ISBN: 978-1-61208-203-5

where $[\boldsymbol{r}]_j$ $(j = 0, \ldots, K-2)$ denotes the jth element of $\boldsymbol{r}$, and s_0 and s_1 are given by (4). Note here that since we assume that $\{s^{(i)}(t)\}$ is stationary for $i = 3, \ldots, K$, the stationary probability vector $\boldsymbol{r}$ is also the initial state probability vector of the Markov chain $\{s^{(i)}(t)\}$ for $i = 3, \ldots, K$.

We next define a $4(K-1) \times 4(K-1)$ matrix $\boldsymbol{Q}$ by

$$\boldsymbol{Q} = \boldsymbol{P} \otimes \boldsymbol{P} \otimes \boldsymbol{R}, \tag{9}$$

where $\otimes$ denotes the Kronecker product, $\boldsymbol{P}$ is determined by (5) and (6), and $\boldsymbol{R}$ is defined by (7). Note that the matrix $\boldsymbol{Q}$ is a transition probability matrix for the Markov chains $\{s^{(i)}(t)\}$ for $i = 1, \ldots, K$.

Let $\psi(k, n, x)$ denote the probability that given that k MSs is in state 1, the number of minislots is equal to n and the feedback probability is equal to x, the k MSs fail to feed back. For $k = 0, \ldots, K-2$, $n = 1, 2, \ldots$ and $0 \le x \le 1$, $\psi(k, n, x)$ is given by

$$\psi(k, n, x) = [1 - k(1-x)^{k-1}x]^n.$$

We then define a $4(K-1) \times 4(K-1)$ diagonal matrix $\boldsymbol{D}(z)$ by

$$\boldsymbol{D}(z) = \mathrm{diag}(\boldsymbol{d}_{0,0}(z), \boldsymbol{d}_{0,1}(z), \boldsymbol{d}_{1,0}(z), \boldsymbol{d}_{1,1}(z)), \tag{10}$$

where $\boldsymbol{d}_{i,j}(z)$ $(i, j = 0, 1)$ is a $1 \times (K-1)$ vector given by

$$\begin{aligned}
[\boldsymbol{d}_{0,0}(z)]_k &= \psi(k, N, u)\frac{z + z^{-1} + K - 2}{K} \\
&\quad + 1 - \psi(k, N, u), \\
[\boldsymbol{d}_{0,1}(z)]_k &= \psi(k+1, N, u)\frac{z + z^{-1} + K - 2}{K} \\
&\quad + (1 - \psi(k+1, N, u))\frac{z^{-1} + k}{k+1}, \\
[\boldsymbol{d}_{1,0}(z)]_k &= \psi(k+1, N, u)\frac{z + z^{-1} + K - 2}{K} \\
&\quad + (1 - \psi(k+1, N, u))\frac{z + k}{k+1}, \\
[\boldsymbol{d}_{1,1}(z)]_k &= \psi(k+2, N, u)\frac{z + z^{-1} + K - 2}{K} \\
&\quad + (1 - \psi(k+2, N, u))\frac{z + z^{-1} + k}{k+2},
\end{aligned}$$

for $k = 0, \ldots, K-2$. We further define $4(K-1) \times 4(K-1)$ matrix $\boldsymbol{C}(z)$ by

$$\boldsymbol{C}(z) = \boldsymbol{D}(z)\boldsymbol{Q}, \tag{11}$$

where $\boldsymbol{D}(z)$ and $\boldsymbol{Q}$ are defined by (10) and (9), respectively. Finally, we define $\eta_n(z)$ $(n = 1, 2, \ldots)$ by

$$\eta_n(z) = (\boldsymbol{r}^{(1)} \otimes \boldsymbol{r}^{(2)} \otimes \boldsymbol{r})\boldsymbol{C}(z)^n \boldsymbol{e},$$

where $\boldsymbol{r}^{(i)}$ denotes the initial state probability vector of the Markov chain $\{s^{(i)}(t)\}$ for $i = 1, 2$, respectively, $\boldsymbol{r}$ denotes the initial state probability vector of the Markov chain for MS i $(i = 3, \ldots, K)$, which is given by (8), $\boldsymbol{e}$ denotes a $4(K-1) \times 1$ vector whose elements are all equal to one, and $\boldsymbol{C}(z)$ is defined by (11).

We are now ready to present the analysis of the STAFI $G_n(x)$. Note that $\eta_n(z)$ can also be expressed in the power series of z as $\eta_n(z) = \sum_{l=-n}^{n} c_l z^l$, where c_l $(l = -n, \ldots, n)$ is a (unknown) real constant satisfying $0 \le c_l \le 1$ and $\sum_{l=-n}^{n} c_l = 1$. Then the probability mass function $g_n(x)$ is expressed as $g_n(x) = c_x + c_{-x}$. Thus, if we determine the unknown real constants $\{c_l\}_{l=-n}^{n}$, we obtain the probability mass function $g_n(x)$. The STAFI $G_n(x)$ is then given by $G_n(x) = \sum_{l=x}^{n} g_n(l) = 1 - \sum_{l=0}^{x-1} g_n(l)$.

There are several possible methods to determine the unknown real constants $\{c_l\}_{l=-n}^{n}$. In this paper, we use the *inverse discrete FFT method* [18] to determine them. Since $g_n(x)$ has a finite support, i.e., $g_n(x) = 0$ for $x > n$, we can calculate the exact value of $g_n(x)$ by using the inverse discrete FFT method.

For comparison, we consider a random scheduler which randomly selects a MS among K MSs irrespective of their received SNRs. For the STAFI of the random scheduler, we define $\tilde{\eta}_n(z)$ $(n = 1, 2, \ldots)$ by

$$\tilde{\eta}_n(z) = \left(\frac{z + z^{-1} + K - 2}{K}\right)^n,$$

which corresponds to $\eta_n(z)$ of the opportunistic feedback fair scheduler. Similar to the case of the opportunistic feedback fair scheduler, from $\tilde{\eta}_n(z)$, we can calculate the exact value the STAFI $\tilde{G}_n(x)$ and the probability mass function $\tilde{g}_n(x)$ for the random scheduler.

IV. Numerical results

In this section, we provide numerical results to investigate the properties of the STAFI of the opportunistic feedback fair scheduler. Throughout numerical results provided in this subsection, we set the parameters as $f_d = 10$ Hz and $T_f = 1$ msec where we decided these parameter values according to [15], [19]. In the numerical results provided in this paper, we also set the initial state probability vectors $\boldsymbol{r}^{(1)}$ and $\boldsymbol{r}^{(2)}$ of MS 1 and MS 2 to the stationary probability vector $\boldsymbol{s}$.

First, we observe the effect of the threshold γ_1 on the STAFI $G_n(x)$. Figure 2 displays the STAFI $G_{256}(x)$ of the opportunistic feedback fair scheduler as a function of x. In Figure 2, we set the number of MSs K, the number of minislots N and the feedback probability u to 30, 5 and 0.80, respectively. For comparison, Figure 2 also shows the STAFI $G_{256}(x)$ of the random scheduler. In the figures, "OFF(x dB)" means the opportunistic feedback fair scheduler whose threshold γ_1 is equal to x, and "RS" means the random scheduler.

In Figure 2, we observe the following. For whole range of x, the STAFIs $G_{256}(x)$ of the opportunistic feedback fair schedulers are greater than the STAFI $G_{256}(x)$ of the random scheduler. In other words, the short term fairness provided by the opportunistic feedback fair schedulers is worse than that provided by the random scheduler. This is due to the positive correlation of the normalized SNR process $\{z^{(i)}(t)/\bar{z}^{(i)}\}$ in

 ISBN: 978-1-61208-203-5

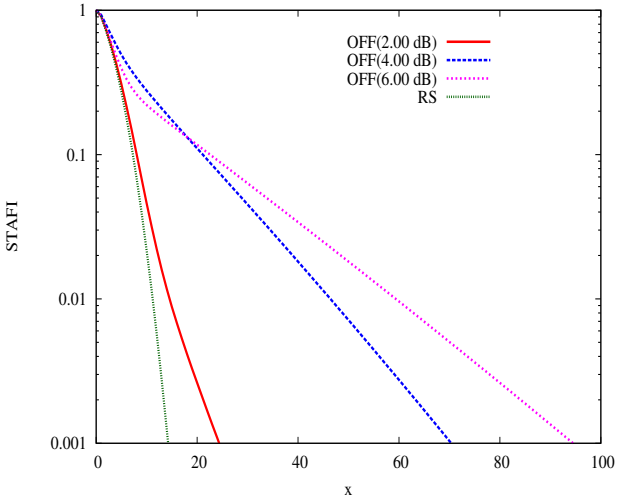


Fig. 2. Effect of γ_1 on STAFI $G_{256}(x)$ $(u = 0.80)$

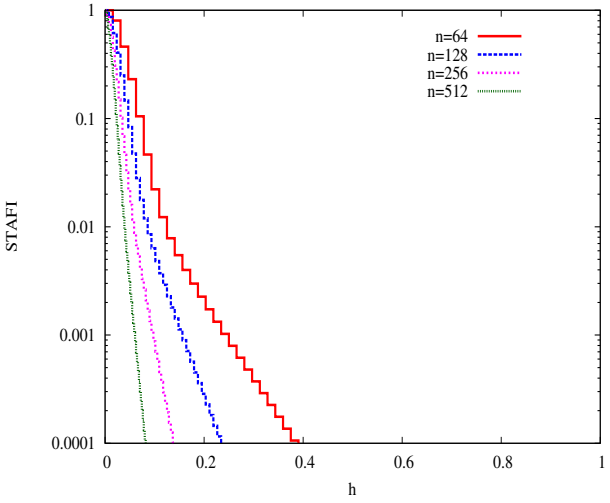


Fig. 4. STAFI $G_n(hn)$ as a function of h ($\gamma_1 = 2.00$ dB)

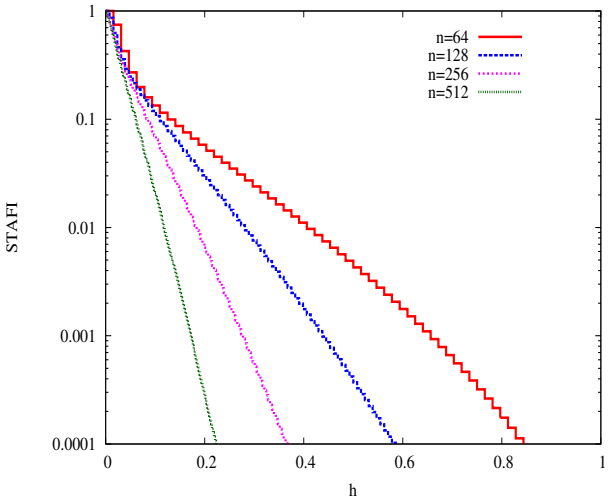


Fig. 3. STAFI $G_n(hn)$ as a function of h ($\gamma_1 = 4.00$ dB)

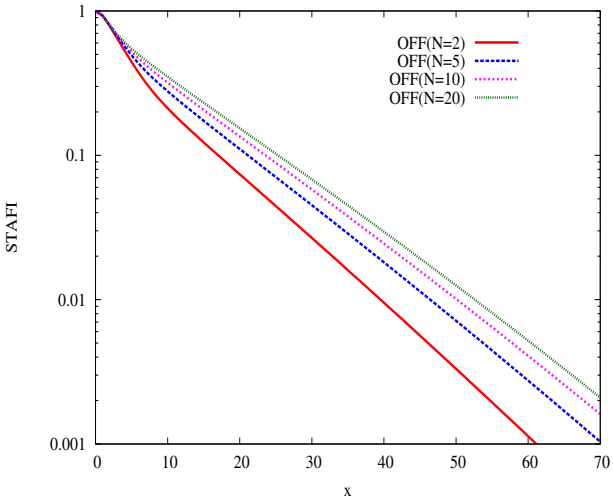


Fig. 5. Effect of number of minislots N on STAFI $G_{256}(x)$

time. We also see that for whole range of x, the OFF (2.00 dB) yields better short term fairness than the OFF (4.00 dB) and the OFF (6.00dB). Comparing the OFF (4.00 dB) and the OFF (6.00dB), we observe that for small x of $G_{256}(x)$, the OFF (6.00 dB) provides better fairness than the OFF (4.00 dB) However, the situation is converse for large x of $G_{256}(x)$. Thus, the OFF (6.00 dB) can keep the probability of moderate unfairness lower, but it can cause serious unfairness with higher probability, compared to the OFF (4.00 dB). A similar non-monotonous property about the threshold value has been observed for the one-bit feedback fair scheduler, too [12].

We next examine how the STAFI of the opportunistic feedback fair scheduler changes as the increase of observation period n. Figures 3 and 4 exhibit the STAFI $G_n(hn)$ as a function of h for $n = 64, 128, 256, 512$. In the figures, we set the number of MSs K, the number of minislots N and the feedback probability u to 30, 5, 0.8, respectively. We set the threshold γ_1 to 4.00 dB in Figure 3 and to 2.00 dB in Figure 4. In Figures 3 and 4, we observe that the STAFI $G_n(hn)$ of the opportunistic feedback fair scheduler rapidly decreases with increase of the observation period n for every h. In other words, the STAFI of the opportunistic feedback fair scheduler rapidly approaches to the ideal long term fairness as the progress of time.

Next, we observe the effect of the number of minislots N on the STAFI $G_n(x)$. Figure 5 displays the STAFI $G_{256}(x)$ of the opportunistic feedback fair scheduler as a function of x. In Figure 5, we set the number of MSs K, the threshold γ_1 and the feedback probability u to 30, 4.00 dB and 0.80, respectively. In the figures, "OFF(N=x)" means the opportunistic feedback fair scheduler where the number of minislots N is equal to x. In Figure 5, we observe that with the increase in the number of minislots N, the short term fairness of the

 ISBN: 978-1-61208-203-5

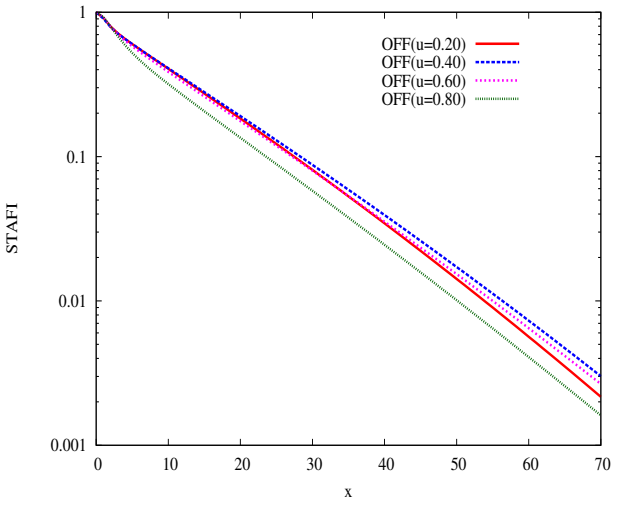


Fig. 6. Effect of feedback probability u on STAFI $G_{256}(x)$

opportunistic feedback fair schedulers becomes worse.

Finally, we observe the effect of the feedback probability u on the STAFI $G_n(x)$. Figure 6 displays the STAFI $G_{256}(x)$ of the opportunistic feedback fair scheduler as a function of x. In Figure 6, we set the number of MSs K, the threshold γ_1 and the number of minislots N to 30, 4.00 dB and 10, respectively. In the figure, "OFF(u=x)" means the opportunistic feedback fair scheduler where the feedback probability is equal to x. In Figure 6, we observe that among the four opportunistic feedback fair schedulers for $u = 0.2, 0.4, 0.6, 0.8$, the scheduler for $u = 0.4$ yields the worst short term fairness. When the feedback probability u is small, the short term fairness of the opportunistic feedback fair scheduler becomes worse with the increase in the feedback probability u. However, if the feedback probability is greater than a certain value, the short term fairness becomes better with the increase in the feedback probability.

V. Conclusion

In this paper, considering the STAFI as a measure of short term fairness, we studied the short term fairness provided by the opportunistic feedback fair scheduler. We developed a numerical method to to calculate the exact value of the STAFI by using the inverse discrete FFT method. In the numerical results, we observed that the threshold γ_1 strongly affects the properties of the short term fairness provided by the opportunistic feedback fair scheduler. The opportunistic feedback fair scheduler with larger threshold γ_1 can keep the probability of moderate unfairness lower, but is can cause serious unfairness with higher probability, compared to the opportunistic feedback fair scheduler with smaller threshold. The impacts of the number of minislots N and the feedback probability u on the properties of short term fairness do not seem to be so strong, compared to the effect of the threshold γ_1. We also observed that the STAFI of the opportunistic feedback fair scheduler approaches to the ideal fairness in a relatively short time period. However, if rigorous fairness is required even in a relatively short time period, we should carefully determine the threshold value γ_1 by considering the short term fairness of the scheduler as well as its information theoretic capacity.

References

[1] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," *Proc. of IEEE ICC '95*, pp. 331–335, 1995.

[2] F. Florén, O. Edfors and B.-A. Molin, "The effect of feedback quantization on the throughput of a multiuser diversity scheme," *Proc. of IEEE GLOBECOM 2003*, pp. 497-501, 2003.

[3] F. Ishizaki and G. U. Hwang, "Queuing delay analysis for packet schedulers with/without multiuser diversity over a fading channel," *IEEE Trans. Veh. Technol.*, vol. 56, no. 5, pp. 3220–3227, 2007.

[4] J. So and J. M. Cioffi, "Feedback reduction scheme for downlink multiuser diversity," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 668–672, 2009.

[5] J. W. So and J. M. Cioffi, "Capacity and fairness in multiuser diversity systems with opportunistic feedback," *IEEE Communications Letters*, vol. 12, no. 9, pp. 648–650, 2008.

[6] H. Kim and Y. Han, "An opportunistic channel quality feedback scheme for proportional fair scheduling," *IEEE Communications Letters*, vol. 11, no. 6, pp. 501–503, 2007.

[7] T. Tang and R. W. Heath Jr., "Opportunistic feedback for downlink multiuser diversity," *IEEE Communications Letters*, vol. 9, no. 10, pp. 948–950, 2005.

[8] D. Wu and R. Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *IEEE Trans. Veh. Technol.*, vol. 54, no. 3, pp. 1198-1206, 2005.

[9] L. Yang, M. Kang, and M.-S. Alouini, "On the capacity-fairness tradeoff in multiuser diversity systems," *IEEE Trans. Veh. Technol.*, vol. 56, no. 4, pp. 1901–1907, 2007.

[10] J. Diaz, O. Simeone, and Y. Bar-Ness, "Sum-rate of MIMO broadcast channels with one bit feedback," *Proc. of IEEE International Symposium on Information Theory (ISIT '06)*, pp. 1944–1948, 2006.

[11] G. U. Hwang and F. Ishizaki, "Design of a fair scheduling exploiting multiuser diversity with feedback reduction," *IEEE Communications Letters*, vol. 12, no. 2, pp. 124-126, 2008.

[12] F. Ishizaki, "Analysis of the statistical time-access fairness index of one-bit feedback fair scheduler," *Numerical Algebra Control and Optimization*, vol. 1, no. 4, pp. 675–689, 2011.

[13] O. Somekh, A.M. Haimovich, and Y. Bar-Ness, "Sum-rate analysis of downlink channels with 1-bit feedback," *IEEE Communications Letters*, vol. 11, no. 2, pp. 137–139, 2007.

[14] D. Gesbert and M.-S. Alouini, "How much feedback is multi-user diversity really worth?," *Proc. of IEEE ICC '04*, pp. 234–238, 2004.

[15] G. U. Hwang and F. Ishizaki, "Analysis of short term fairness and its impact on packet level performance," *Performance Evaluation*, vol. 67, no. 12, pp. 1340-1352, 2010.

[16] B. Tan, L. Ying, and R. Srikant, "Short-term fairness and long-term QoS," *Proc. of Conference on Information Science and Systems (CISS)*, pp. 1201–1204, 2008.

[17] Y. Liu, S. Gruhl, and E. W. Knightly, "WCFQ: an opportunistic wireless scheduler with statistical fairness bounds," *IEEE Trans. Wireless Commun.*, vol. 2, no. 5, pp. 1017–1028, 2003.

[18] H. C. Tijms, *A first course in stochastic models*, John Wiley & Sons, 2003.

[19] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 1142–1153, 2005.

[20] M. D. Yacoub, *Foundation of mobile radio engineering*, Boca Ration, FL:CRC, 1993.

 ISBN: 978-1-61208-203-5

Probability Density Functions of Derivatives in Two Time Instants for SSC Combiner in Rician Fading Channel

Dragana Krstić
Department of Telecommunications,
Faculty of Electronic Engineering, University of Niš
Aleksandra Medvedeva 14
Niš, Serbia
dragana.krstic@elfak.ni.ac.rs

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

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

***Abstract*—The probability density functions (PDFs) of derivatives in two time instants for output signals from dual branch Switch and Stay Combiner (SSC) in the presence of Rician fading are determined in this paper. The second order statistics such as the average level crossing rate and the average fade duration can be calculated by using obtained closed-form expressions.**

Keywords-probability density function; Rician fading; Switch and Stay Combining; time derivative

I. INTRODUCTION

Fading is one of the most important causes of degradation signals in wireless communication systems [1]. Ricean fading is a stochastic model for radio propagation anomaly caused by partial cancellation of a radio signal by itself — the signal arrives at the receiver by several different paths (hence exhibiting multipath interference), and at least one of the paths is changing (lengthening or shortening). Rician fading occurs when one of the paths, typically a line of sight signal, is much stronger than the others. In Rician fading, the amplitude gain is characterized by a Rician distribution [2], [3].

Rayleigh fading is the specialized model for stochastic fading when there is no line of sight signal, and is sometimes considered as a special case of the more generalized concept of Rician fading. In Rayleigh fading, the amplitude gain is characterized by a Rayleigh distribution.

In telecommunications, a diversity scheme refers to a method for improving the reliability of a message signal by using two or more communication channels with different characteristics. Diversity plays an important role in combating fading effect and co-channel interference and avoiding errors [4]-[6]. It is based on the fact that individual channels experience different levels of fading and interference. Multiple versions of the same signal may be transmitted or received and combined in the receiver. Diversity techniques may exploit the multipath propagation, resulting in a diversity gain, often measured in decibels.

When space diversity is used the signal is transmitted over several different propagation paths. In the case of wired transmission, this can be achieved by transmitting via multiple wires. In the case of wireless transmission, it can be achieved by antenna diversity using multiple transmitter antennas (transmit diversity) and/or multiple receiving antennas (reception diversity). In the latter case, a diversity combining technique is applied before further signal processing takes place.

Diversity combining is the technique applied to combine the multiple received signals of a diversity reception device into a single improved signal. Various diversity combining techniques can be distinguished:

Selection combining (SC): Of the N received signals, the strongest signal is selected [7]. When the N signals are independent and Rayleigh distributed, the expected diversity gain has been shown to be inversely proportional to the number of antennas [8, 9]. Therefore, any additional gain diminishes rapidly with the increasing number of channels.

Switched combining: The receiver switches to another signal when the currently selected signal drops below a predefined threshold [10, 11]. This is a less efficient technique than selection combining.

Equal-gain combining (EGC): All the received signals are summed coherently [12].

Maximal-ratio combining (MRC) is often used in large phased-array systems. The received signals are weighted with respect to their SNR and then summed [13].

The authors determined earlier the probability density functions and joint probability density functions for SSC combiner output signals at two time instants in the presence of different fading distributions and used these expressions for obtaining better system performances, such as the bit error rate and the outage probability, for complex systems sampling at two time instants. Performance analysis of SSC/SC combiner in the presence of Rayleigh and log-normal fading are performed in [14] and [15], respectively.

In this paper, the probability density functions (PDFs) of derivatives for Switch and Stay Combiner (SSC) output signals at two time instants in the presence of Rician fading will be determined. The dual branch SSC combiner will be

 ISBN: 978-1-61208-203-5

considered. Subsequently, the second-order characteristics can be determined using these PDF [16].

The remainder of the document is organized in the following way: Section II introduces the model of the SSC combiner observed and basic assumptions of the problem under consideration. After that, in Section III, the probability density function of derivative is derived and graphicaly presented. Last section gives some conclusions.

II. SYSTEM MODEL

This section discusses the SSC combiner with two branches in two time moments. The model is shown in Figure 1. The input signals are r_{11} and r_{21} in the first time moment, and r_{12} and r_{22} in the second time moment. The signals at the output are r_1 and r_2. The derivatives are $\dot{r}_{11}$ and $\dot{r}_{21}$ at the first time moment, and $\dot{r}_{12}$ and $\dot{r}_{22}$ at the second time moment. The derivatives at the SSC combiner output are $\dot{r}_1$ and $\dot{r}_2$.

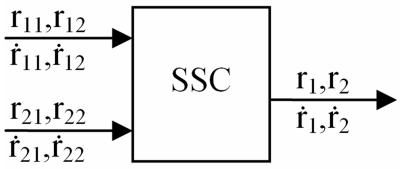


Figure 2. Model of the SSC combiner with two inputs at two time instants

The indices for input signals and their derivatives are: the first index represents the branch ordinal number and the other one signs the time instant observed. The indices for the output signal correspond to the time instants considered.

The probability that combiner examines first the signal from the first branch is P_1 and P_2 for the second. The values of P_1 and P_2 for SSC combiner are obtained in [1].

The four different cases are discussed here:

1) $r_1<r_T,\ r_2<r_T$

In this case all signals are less then threshold r_T, i.e.: $r_{11}<r_T,\ r_{12}<r_T\ \ r_{21}<r_T$, and $r_{22}<r_T$. Let combiner considers first the signal r_{11}. Because $r_{11}<r_T$, then $\dot{r}_1=\dot{r}_{21}$, and because of $r_{22}<r_T$, then $\dot{r}_2=\dot{r}_{12}$. The probability of this event is P_1. If combiner examines first the signal r_{21}, then $r_{21}<r_T,\ \dot{r}_1=\dot{r}_{11}$, as $r_{21}<r_T,\ \dot{r}_2=\dot{r}_{22}$. The probability of this event is P_2.

2) $r_1\geq r_T,\ r_2<r_T$

The possible combinations are:

- $r_{11}\geq r_T,\quad r_{12}<r_T,\ r_{22}<r_T,\qquad \dot{r}_1=\dot{r}_{11}\quad \dot{r}_2=\dot{r}_{22}\quad P_1$

- $r_{11}<r_T,\ r_{21}\geq r_T\quad r_{22}<r_T,\ r_{12}<r_T,\qquad \dot{r}_1=\dot{r}_{21}\quad \dot{r}_2=\dot{r}_{12}\quad P_1$

- $r_{21}\geq r_T,\quad r_{22}<r_T,\ r_{12}<r_T,\qquad \dot{r}_1=\dot{r}_{21}\quad \dot{r}_2=\dot{r}_{12}\quad P_2$

- $r_{21}<r_T,\ r_{11}\geq r_T,\quad r_{12}<r_T,\ r_{22}<r_T,\qquad \dot{r}_1=\dot{r}_{11}\quad \dot{r}_2=\dot{r}_{22}\quad P_2$

3) $r_1<r_T,\ r_2\geq r_T$

The possible combinations for this case are:

- $r_{11}<r_T,\ r_{21}<r_T,\quad r_{22}\geq r_T,\qquad \dot{r}_1=\dot{r}_{21}\quad \dot{r}_2=\dot{r}_{22}\quad P_1$

- $r_{11}<r_T,\ r_{21}<r_T,\quad r_{22}<r_T,\ r_{12}\geq r_T,\qquad \dot{r}_1=\dot{r}_{21}\quad \dot{r}_2=\dot{r}_{12}\quad P_1$

- $r_{21}<r_T,\ r_{11}<r_T,\quad r_{12}\geq r_T,\qquad \dot{r}_1=\dot{r}_{11}\quad \dot{r}_2=\dot{r}_{12}\quad P_2$

- $r_{21}<r_T,\ r_{11}<r_T,\quad r_{12}<r_T,\ r_{22}\geq r_T,\qquad \dot{r}_1=\dot{r}_{11}\quad \dot{r}_2=\dot{r}_{22}\quad P_2$

4) $r_1\geq r_T,\ r_2\geq r_T$

Now, the possible combinations are:

- $r_{11}\geq r_T,\ r_{12}\geq r_T,\qquad \dot{r}_1=\dot{r}_{11}\quad \dot{r}_2=\dot{r}_{12}\quad P_1$

- $r_{11}\geq r_T,\ r_{12}<r_T,\ r_{22}\geq r_T\qquad \dot{r}_1=\dot{r}_{11}\quad \dot{r}_2=\dot{r}_{22}\quad P_1$

- $r_{11}<r_T,\ r_{21}\geq r_T,\ r_{22}\geq r_T,\qquad \dot{r}_1=\dot{r}_{21}\quad \dot{r}_2=\dot{r}_{22}\quad P_1$

- $r_{11}<r_T,\ r_{21}\geq r_T,\ r_{22}<r_T,\ r_{12}<r_T\qquad \dot{r}_1=\dot{r}_{21}\quad \dot{r}_2=\dot{r}_{12}\quad P_1$

- $r_{21}\geq r_T,\ r_{22}\geq r_T,\qquad \dot{r}_1=\dot{r}_{21}\quad \dot{r}_2=\dot{r}_{22}\quad P_2$

- $r_{21}\geq r_T,\ r_{22}<r_T,\ r_{12}\geq r_T,\qquad \dot{r}_1=\dot{r}_{21}\quad \dot{r}_2=\dot{r}_{12}\quad P_2$

- $r_{21}<r_T,\ r_{11}\geq r_T,\ r_{12}\geq r_T,\qquad \dot{r}_1=\dot{r}_{11}\quad \dot{r}_2=\dot{r}_{12}\quad P_2$

- $r_{21}<r_T,\ r_{11}\geq r_T,\quad r_{12}<r_T,\ r_{22}\geq r_T,\qquad \dot{r}_1=\dot{r}_{11}\quad \dot{r}_2=\dot{r}_{22}\quad P_2$

III. PROBABILITY DENSITY FUNCTIONS OF DERIVATIVES

The joint probability density functions of signal derivatives are:

$r_1<r_T,\ r_2<r_T$

$$p_{r_1r_2\dot{r}_1\dot{r}_2}(r_1,r_2,\dot{r}_1,\dot{r}_2)=P_1\int_0^{r_T}dr_{11}\int_0^{r_T}dr_{22}\,p_{r_{11}r_{22}r_{21}r_{12}\dot{r}_{21}\dot{r}_{12}}(r_{11},r_{22},r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_2\int_0^{r_T}dr_{21}\int_0^{r_T}dr_{12}\,p_{r_{21}r_{12}r_{11}r_{22}\dot{r}_{11}\dot{r}_{22}}(r_{21},r_{12},r_1,r_2,\dot{r}_1,\dot{r}_2) \qquad (1)$$

$r_1\geq r_T,\ r_2<r_T$

$$p_{r_1r_2\dot{r}_1\dot{r}_2}(r_1,r_2,\dot{r}_1,\dot{r}_2)=P_1\int_0^{r_T}dr_{12}\,p_{r_{12}r_{11}r_{22}\dot{r}_{11}\dot{r}_{22}}(r_{12},r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_1\int_0^{r_T}dr_{11}\int_0^{r_T}dr_{22}\,p_{r_{11}r_{22}r_{21}r_{12}\dot{r}_{21}\dot{r}_{12}}(r_{11},r_{22},r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_2\int_0^{r_T}dr_{22}\,p_{r_{22}r_{21}r_{12}\dot{r}_{21}\dot{r}_{12}}(r_{22},r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_2\int_0^{r_T}dr_{21}\int_0^{r_T}dr_{12}\,p_{r_{21}r_{12}r_{11}r_{22}\dot{r}_{11}\dot{r}_{22}}(r_{21},r_{12},r_1,r,\dot{r}_1,\dot{r}_2) \quad (2)$$

$r_1<r_T,\ r_2\geq r_T$

$$p_{r_1r_2\dot{r}_1\dot{r}_2}(r_1,r_2,\dot{r}_1,\dot{r}_2)=P_1\int_0^{r_T}dr_{11}\,p_{r_{11}r_{21}r_{22}\dot{r}_{21}\dot{r}_{22}}(r_{11},r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_1\int_0^{r_T}dr_{11}\int_0^{r_T}dr_{22}\,p_{r_{11}r_{22}r_{21}r_{12}\dot{r}_{21}\dot{r}_{12}}(r_{11},r_{22},r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_2\int_0^{r_T}dr_{21}\,p_{r_{21}r_{11}r_{12}\dot{r}_{11}\dot{r}_{12}}(r_{21},r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_2\int_0^{r_T}dr_{21}\int_0^{r_T}dr_{12}\,p_{r_{21}r_{12}r_{11}r_{22}\dot{r}_{11}\dot{r}_{22}}(r_{21},r_{12},r_1,r,\dot{r}_1,\dot{r}_2) \quad (3)$$

$r_1\geq r_T,\ r_2\geq r_T$

$$p_{r_1r_2\dot{r}_1\dot{r}_2}(r_1,r_2,\dot{r}_1,\dot{r}_2)=P_1p_{r_{11}r_{12}\dot{r}_{11}\dot{r}_{12}}(r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_1\int_0^{r_T}dr_{12}\,p_{r_{12}r_{11}r_{22}\dot{r}_{11}\dot{r}_{22}}(r_{12},r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_1\int_0^{r_T}dr_{11}\,p_{r_{11}r_{21}r_{22}\dot{r}_{21}\dot{r}_{22}}(r_{11},r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_1\int_0^{r_T}dr_{11}\int_0^{r_T}dr_{22}\,p_{r_{11}r_{22}r_{21}r_{12}\dot{r}_{21}\dot{r}_{12}}(r_{11},r_{22},r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_2p_{r_{21}r_{22}\dot{r}_{21}\dot{r}_{22}}(r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_2\int_0^{r_T}dr_{22}\,p_{r_{22}r_{21}r_{12}\dot{r}_{21}\dot{r}_{12}}(r_{22},r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_2\int_0^{r_T}dr_{21}\,p_{r_{21}r_{11}r_{12}\dot{r}_{11}\dot{r}_{12}}(r_{21},r_1,r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_2\int_0^{r_T}dr_{21}\int_0^{r_T}dr_{12}\,p_{r_{21}r_{12}r_{11}r_{22}\dot{r}_{11}\dot{r}_{22}}(r_{21},r_{12},r_1,r,\dot{r}_1,\dot{r}_2) \quad (4)$$

For the case that signal and its derivative are not correlated, after integrating of the whole range of signal values and some mathematical manipulations, the joint PDF of derivative can be expressed as:

$$p_{\dot{r}_1\dot{r}_2}(\dot{r}_1,\dot{r}_2)=P_1\int_0^{r_T}dr_{11}\int_0^{r_T}dr_{22}\,p_{r_{11}r_{22}\dot{r}_{21}\dot{r}_{12}}(r_{11},r_{22},\dot{r}_1,\dot{r}_2)+$$

$$+P_2\int_0^{r_T}dr_{21}\int_0^{r_T}dr_{12}\,p_{r_{21}r_{12}\dot{r}_{11}\dot{r}_{22}}(r_{21},r_{12},\dot{r}_1,\dot{r}_2)+$$

$$+P_1\int_0^{r_T}dr_{12}\int_{r_T}^{\infty}dr_1\,p_{r_{12}r_{11}\dot{r}_{11}\dot{r}_{22}}(r_{12},r_1,\dot{r}_1,\dot{r}_2)+P_2\int_0^{r_T}dr_{22}\int_{r_T}^{\infty}dr_1\,p_{r_{22}r_{21}\dot{r}_{21}\dot{r}_{12}}(r_{22},r_1,\dot{r}_1,\dot{r}_2)+$$

$$+P_1\int_0^{r_T}dr_{11}\int_{r_T}^{\infty}dr_2\,p_{r_{11}r_{22}\dot{r}_{21}\dot{r}_{22}}(r_{11},r_2,\dot{r}_1,\dot{r}_2)+P_2\int_0^{r_T}dr_{21}\int_{r_T}^{\infty}dr_2\,p_{r_{21}r_{12}\dot{r}_{11}\dot{r}_{12}}(r_{21},r_2,\dot{r}_1,\dot{r}_2)+$$

$$+P_1\int_{r_T}^{\infty}dr_1\int_{r_T}^{\infty}dr_2\,p_{r_{11}r_{12}\dot{r}_{11}\dot{r}_{12}}(r_1,r_2,\dot{r}_1,\dot{r}_2)+P_2\int_{r_T}^{\infty}dr_1\int_{r_T}^{\infty}dr_2\,p_{r_{21}r_{22}\dot{r}_{21}\dot{r}_{22}}(r_1,r_2,\dot{r}_1,\dot{r}_2) \quad (5)$$

The signal derivatives PDFs can be found from joint PDF based on:

$$p_{\dot{r}_1}(\dot{r}_1)=\int_{-\infty}^{\infty}p_{\dot{r}_1\dot{r}_2}(\dot{r}_1,\dot{r}_2)d\dot{r}_2 \quad (6)$$

$$p_{\dot{r}_2}(\dot{r}_2)=\int_{-\infty}^{\infty}p_{\dot{r}_1\dot{r}_2}(\dot{r}_1,\dot{r}_2)d\dot{r}_1 \quad (7)$$

By replacing (5) in (6) and (7), obtained:

$$p_{\dot{r}_1}(\dot{r}_1)=P_1p_{\dot{r}_{11}}(\dot{r}_1)+P_2p_{\dot{r}_{21}}(\dot{r}_1)+$$

$$+\left(P_2F_{r_{21}}(r_T)-P_1F_{r_{11}}(r_T)\right)p_{\dot{r}_{11}}(\dot{r}_1)+$$

$$+\left(P_1F_{r_{11}}(r_T)-P_2F_{r_{21}}(r_T)\right)p_{\dot{r}_{21}}(\dot{r}_1) \quad (8)$$

$$p_{\dot{r}_2}(\dot{r}_2)=P_1F_{r_{11}}(r_T)F_{r_{22}}(r_T)p_{\dot{r}_{12}}(\dot{r}_2)+P_2F_{r_{21}}(r_T)F_{r_{12}}(r_T)p_{\dot{r}_{22}}(\dot{r}_2)+$$

$$+P_1B_1(r_T)p_{\dot{r}_{22}}(\dot{r}_2)+P_2B_2(r_T)p_{\dot{r}_{12}}(\dot{r}_2)+$$

$$+P_1F_{r_{11}}(r_T)\left(1-F_{r_{22}}(r_T)\right)p_{\dot{r}_{22}}(\dot{r}_2)+P_2F_{r_{21}}(r_T)\left(1-F_{r_{12}}(r_T)\right)p_{\dot{r}_{12}}(\dot{r}_2)+$$

$$+P_1C_1(r_T)p_{\dot{r}_{12}}(\dot{r}_2)+P_2C_2(r_T)p_{\dot{r}_{22}}(\dot{r}_2) \quad (9)$$

where $F_{r_{ij}}(r_T)$ are signals' CDFs and $F_{r_{i1}}(r_T)=F_{r_{i2}}(r_T)$, while $B_i(r_T)$ and $C_i(r_T)$ are obtained based on [(11), 17]

$$B_i(r_T)=\left(1-\rho_i^{\,2}\right)e^{-\frac{A_i^2}{\sigma_i^2(1+\rho_i)}}\cdot$$

$$\cdot\sum_{k,l_1,l_2,l_3=0}^{\infty}\varepsilon_k\cdot\frac{1}{l_1!l_2!l_3!(k+l_1)!(k+l_2)!(k+l_3)!}\cdot$$

$$\cdot\rho^{k+2l_1}\left[\left(\frac{1-\rho_i}{1+\rho_i}\right)\left(\frac{A_i}{2\sigma_i^{\,2}}\right)\right]^{k+l_2+l_3}\cdot$$

$$\cdot\gamma\left(k+l_1+l_2+1,\frac{r_T^{\,2}}{2\sigma_i^{\,2}(1-\rho_i^{\,2})}\right)\cdot$$

$$\cdot\left[1-\gamma\left(k+l_1+l_3+1,\frac{r_T^{\,2}}{2\sigma_i^{\,2}(1-\rho_i^{\,2})}\right)\right] \quad (10)$$

$$C_i(r_T)=\left(1-\rho_i^{\,2}\right)e^{-\frac{A_i^2}{\sigma_i^2(1+\rho_i)}}\cdot$$

$$\cdot\sum_{k,l_1,l_2,l_3=0}^{\infty}\varepsilon_k\cdot\frac{1}{l_1!l_2!l_3!(k+l_1)!(k+l_2)!(k+l_3)!}\cdot$$

$$\cdot\rho^{k+2l_1}\left[\left(\frac{1-\rho_i}{1+\rho_i}\right)\left(\frac{A_i}{2\sigma_i^{\,2}}\right)\right]^{k+l_2+l_3}\cdot$$

$$\cdot\left[1-\gamma\left(k+l_1+l_2+1,\frac{r_T^{\,2}}{2\sigma_i^{\,2}(1-\rho_i^{\,2})}\right)\right]\cdot$$

$$\cdot\left[1-\gamma\left(k+l_1+l_3+1,\frac{r_T^{\,2}}{2\sigma_i^{\,2}(1-\rho_i^{\,2})}\right)\right] \quad (11)$$

$\gamma()$ is incomplete gamma function and ε_k is Neumman factor defined by

$$\varepsilon_k=\begin{cases}1, & k=0\\ 2, & k>0\end{cases}$$

The probability density functions of signal derivatives in the presence of Rician fading at the combiner input has normal distribution with zero mean value [18, 19]:

$$p_{\dot r_i}(\dot r_{i,j})=\frac{1}{\sqrt{2\pi}\dot\sigma_i}e^{-\frac{\dot r_{i,j}^{\,2}}{2\dot\sigma_i^{\,2}}} \quad ,-\infty<\dot r_{i,j}<\infty \quad (12)$$

where $i=1,2;\ j=1,2$ and $\dot\sigma_i^{\,2}=2\sigma_i^{\,2}\pi^2 f_m^{\,2}$ is the variance and f_m is maximal Doppler frequency.

Probability density function of signal derivatives $\dot r_1$ and $\dot r_2$ at the SSC combiner output at two time moments in the presence of Rician fading is obtained when (12) putting in previously obtained general expressions for PDFs of signal derivatives and replacing of CDF with [20]:

$$F_{r_i}(r_{i,j})=1-Q_1\left(A_i/\sigma_i,r_{i,j}/\sigma_i\right),\quad r_{i,j}\ge 0 \quad (13)$$

where $Q_1(\ )$ is Marcum Q-function of first order, are obtained as:

$$p_{\dot r_1}(\dot r_1)=P_1\frac{1}{\sqrt{2\pi}\dot\sigma_1}e^{-\frac{\dot r_1^2}{2\dot\sigma_1^{\,2}}}+P_2\frac{1}{\sqrt{2\pi}\dot\sigma_2}e^{-\frac{\dot r_1^2}{2\dot\sigma_2^{\,2}}}+$$

$$+\left(P_2\left[1-Q_1\left(\frac{A_2}{\sigma_2},\frac{r_T}{\sigma_2}\right)\right]-P_1\left[1-Q_1\left(\frac{A_1}{\sigma_1},\frac{r_T}{\sigma_1}\right)\right]\right)\frac{1}{\sqrt{2\pi}\dot\sigma_1}e^{-\frac{\dot r_1^2}{2\dot\sigma_1^{\,2}}}+$$

$$+\left(P_1\left[1-Q_1\left(\frac{A_1}{\sigma_1},\frac{r_T}{\sigma_1}\right)\right]-P_2\left[1-Q_1\left(\frac{A_2}{\sigma_2},\frac{r_T}{\sigma_2}\right)\right]\right)\frac{1}{\sqrt{2\pi}\dot\sigma_2}e^{-\frac{\dot r_1^2}{2\dot\sigma_2^{\,2}}} \quad (14)$$

$$p_{\dot r_2}(\dot r_2)=P_1\left[1-Q_1\left(\frac{A_1}{\sigma_1},\frac{r_T}{\sigma_1}\right)\right]\left[1-Q_1\left(\frac{A_2}{\sigma_2},\frac{r_T}{\sigma_2}\right)\right]\frac{1}{\sqrt{2\pi}\dot\sigma_1}e^{-\frac{\dot r_2^2}{2\dot\sigma_1^{\,2}}}+$$

$$+P_2\left[1-Q_1\left(\frac{A_1}{\sigma_1},\frac{r_T}{\sigma_1}\right)\right]\left[1-Q_1\left(\frac{A_2}{\sigma_2},\frac{r_T}{\sigma_2}\right)\right]\frac{1}{\sqrt{2\pi}\dot\sigma_2}e^{-\frac{\dot r_2^2}{2\dot\sigma_2^{\,2}}}+$$

$$+P_1B_1(r_T)\frac{1}{\sqrt{2\pi}\dot\sigma_2}e^{-\frac{\dot r_2^2}{2\dot\sigma_2^{\,2}}}+P_2B_2(r_T)\frac{1}{\sqrt{2\pi}\dot\sigma_1}e^{-\frac{\dot r_2^2}{2\dot\sigma_1^{\,2}}}+$$

$$+P_1\left[1-Q_1\left(\frac{A_1}{\sigma_1},\frac{r_T}{\sigma_1}\right)\right]Q_1\left(\frac{A_2}{\sigma_2},\frac{r_T}{\sigma_2}\right)\frac{1}{\sqrt{2\pi}\dot\sigma_2}e^{-\frac{\dot r_2^2}{2\dot\sigma_2^{\,2}}}+$$

$$+P_2\left[1-Q_1\left(\frac{A_2}{\sigma_2},\frac{r_T}{\sigma_2}\right)\right]Q_1\left(\frac{A_1}{\sigma_1},\frac{r_T}{\sigma_1}\right)\frac{1}{\sqrt{2\pi}\dot\sigma_1}e^{-\frac{\dot r_2^2}{2\dot\sigma_1^{\,2}}}+$$

$$+P_1C_1(r_T)\frac{1}{\sqrt{2\pi}\dot\sigma_1}e^{-\frac{\dot r_2^2}{2\dot\sigma_1^{\,2}}}+P_2C_2(r_T)\frac{1}{\sqrt{2\pi}\dot\sigma_2}e^{-\frac{\dot r_2^2}{2\dot\sigma_2^{\,2}}} \quad (15)$$

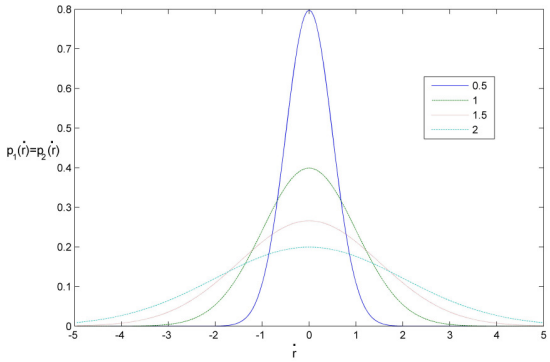


Figure 2. The probability density functions of derivatives at the SSC combiner output at two time instants

The PDFs of signal derivatives are presented in Fig. 2 for different values of parameter $\dot\sigma_i$ in the case of channels with identical distribution.

IV. CONCLUSION

In this paper, the expressions for probability density functions (PDFs) of the time derivatives in two time instants for output signals from dual branch SSC combiner in the presence of Rician fading are obtained. The second order characteristics: the average level crossing rate and the average fade duration for complex combiner who makes the decision based on sampling at two time moments can be

 ISBN: 978-1-61208-203-5

calculated by using closed-form expressions derived in this paper.

ACKNOWLEDGMENT

This work has been funded by the Serbian Ministry for Science under the projects III-44006 and TR-33035.

REFERENCES

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

[2] A. Abdi, C. Tepedelenlioglu, M. Kaveh, and G. Giannakis, "On the estimation of the K parameter for the Rice fading distribution", IEEE Communications Letters, pp. 92 -94, March 2001.

[3] M. A. Richards, Rice Distribution for RCS, Georgia Institute of Technology, Sep. 2006.

[4] J. Moon and Y. Kim. "Antenna Diversity Strengthens Wireless LANs.", Communication Systems Design, pp. 15–22, Jan 2003.

[5] S. M. Lindenmeier, L. M. Reiter, D. E. Barie and J. F. Hopf. "Antenna Diversity for Improving the BER in Mobile Digital Radio Reception Especially in Areas with Dense Foliage." International ITG Conference on Antennas, ISBN 978-3-00-021643-5, pp. 45–48. Mar 30 2007.

[6] C. Dietrich, "Adaptive Arrays and Diversity Antenna Configurations for Handheld Wireless Communication Terminals", Jr. Feb 15, 2000.

[7] D.G. Brennan, "Linear diversity combining techniques," Proc. IRE, vol.47, no.1, pp.1075–1102, June 1959.

[8] D. Milovic, M. Stefanovic, D. Pokrajac, "Stochastic approach for output SINR computation at SC diversity systems with correlated Nakagami-m fading", European Transactions on Telecommunications, vol. 20, no. 5, pp. 482-486, 2009.

[9] A. Cvetković, M. Stefanović, N. Sekulović, D. Milić, D. Stefanović, Z. Popović, "Second-order statistics of dual SC macrodiversity system over channels affected by Nakagami-*m* fading and correlated gamma shadowing", Electrical Review (Przeglad Elektrotechniczny), vol. 87, no. 6, pp. 284-288, June 2011.

[10] P. Spalević, S. Panić, Ć. Dolićanin M. Stefanović, A.Mosić, "SSC Diversiity Receiver over Correlated α-μ Fading Channels in the Presence of co-channel interference", EURASIP Journal on Wireless Communications and Networking, vol. 2010, doi:10.1155/2010/142392.

[11] Đ. V. Banđur, M. Stefanović, M. V. Banđur, "Performance analysis of SSC diversity receiver over correlated Ricean fading channels in the presence of co-channel interference", Electronics Letters, vol. 44, no. 9, pp. 587-588, 2008.

[12] G. T. Djordjevic, D. N. Milic, A. M. Cvetkovic, M. C. Stefanovic, "Influence of Imperfect Cophasing on Performance of EGC Receiver of BPSK and QPSK Signals Transmitted over Weibull Fading Channe", accepted for publication, European Transactions on Telecommunications, published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/ett.1475.

[13] Z. Popovic, S. Panic, J. Anastasov, M. Stefanovic, P.Spalevic, "Cooperative MRC diversity over Hoyt fading channels", Electrical Review, vol. 87 no. 12, pp. 150-152, Dec. 2011.

[14] 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. Vol. 65, Issue 11-12, ISSN (Online) 2191-6349, ISSN (Print) 0016-1136, pp. 319–325, November/2011

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

[16] L. Yang and M. S. Alouini, "Average Level Crossing Rate and Average Outage Duration of Switched Diversity Systems", Global Telecommunications Conference, GLOBECOM '02. IEEE, vol. 2, Print ISBN: 0-7803-7632-3, pp. 1420-1424, 2002.

[17] M. K. Simon, "Comments on Infinite-Series Representations Associated With the Bivariate Rician Distribution and Their Applications", IEEE Trans. Commun., vol. 54, no 8, pp. 2149 – 2153, 25-28 Sept.. 2005.

[18] T. S. Rappaport, Wireless Communications: Principles and Practice. Upper Saddle River, NJ: PTR Prentice-Hall, 1996.

[19] L. Yang, and M.-S. Alouini, "Average Level Crossing Rate and Average Outage Duration of Generalized Selection Combining", IEEE Transactions on Communications, vol. 51, no. 12, pp. 1063-1067, Dec. 2003

[20] S. O. Rice, "Statistical properties of a sine wave plus random noise," Bell Syst. Tech. J., vol. 27, pp. 109–157, Jan. 1948.

 ISBN: 978-1-61208-203-5

Performance Evaluation of MIMO Schemes in 5 MHz Bandwidth LTE System

Ali Jemmali
Departement of Electrical Engineering
Ecole Polytechnique de Montreal
Email: ali.jemmali@polymtl.ca

Jean Conan
Departement of Electrical Engineering
Ecole Polytechnique de Montreal
Email: j.conan@polymtl.ca

***Abstract*—In this paper, we study the performance of the MIMO Schemes in the 3GPP Long Term Evolution (LTE) system with 5 MHz bandwidth. As performances metrics, the Block Error Rate (BLER) and the Data Throughput are evaluated in terms of Signal to Noise Ratio (SNR) for three different Multi-Input Multi-Output (MIMO) schemes as defined in LTE standard. Two transmit diversity schemes known as Space Frequency Block Codes (SFBC) and Frequency Switched Transmit Diversity (FSTD) as well as one Open Loop Spatial Multiplexing (OLSM) scheme are considered in the evaluation. The performance of the three MIMO schemes are compared to the performance of Single Input Single Output (SISO) scheme to evaluate the improvement in BLER and data throughput of the system. The ITU pedestrian B channel with high order modulation and coding scheme is considered for the evaluation.**

Keywords- Multi-antenna MIMO system, LTE, BLER, Data Throughput

I. INTRODUCTION

The 3GPP Long Term Evolution is the latest evolution of the wireless communication systems. LTE is part of the UMTS standards but includes many changes and improvements identified by the 3GPP consortium. The goal of LTE is to increase the data throughput and the speed of wireless data using a combination of new methods and technologies like OFDM and MIMO technics. The LTE downlink transmission is based on Orthogonal Frequency Division Multiple Access (OFDMA). OFDM is a technique of encoding digital data on multiple carrier frequencies and it is known to be efficient to improve the spectral efficiency of wireless system. Another important advantage of OFDM technique is to be more resistant to frequency selective fading than single carrier system by converting the wide-band frequency selective channel into a set of many flat fading subchannels. In addition, OFDMA allows for adding frequency domain scheduling to time domain scheduling. In order to optimize the system data throughput and the coverage area for a given transmission power, LTE make use of the Adaptive Modulation and Coding (AMC). In AMC, the transmitter should assign the data rate for each user depending on the channel quality from the serving cell, the interference level from other cells, and the noise level at the receiver. To achieve the target in terms of data throughput and reliability, the LTE standard makes MIMO as its essential core. MIMO was recognized to be a very powerful technique to improve the performance of wireless communication systems. Multiple antenna techniques can be used in two different modes namely the diversity and multiplexing mode. In diversity mode, the same signal is transmitted over multiple antenna and hence the reliability of the system is improved by the diversity gain. In diversity mode, the mapping function of each symbol to which transmit antenna is called Space Time Block Code (STBC). In multiplexing mode, two different spatial streams are sent from two different antennas and hence the data rate is improved. To study the performance of LTE systems a MATLAB based downlink physical layer simulator [1] [2] for Link Level Simulation (LLS) has been developed. A System Level Simulation [3] of the Simulator is also available. The goal of the development of the simulator was to facilitate comparison with work of different research group and it is publicly available for free under academic non-commercial use license [2]. The main features of the simulator are adaptive coding and modulation, MIMO transmission and scheduling. As the simulator includes many physical layer features, it can be used in different application in research [3]. In [4], the simulator was used to study the channel estimation of OFDM systems and the performance evaluation of a fast fading channel estimator was presented. In [5] and [6], a method for calculating the Precoding Matrix Indicator (PMI), the Rank Indicator (RI) and the Channel Quality Indicator (CQI) were studied and analyzed with the simulator.

In this paper, the BLER and the Data Throughput of SISO and MIMO schemes in 5 MHz LTE system for high Modulation and Coding Scheme (MCS) are investigated in terms of SNR using the Link Level LTE simulator [1] [2]. The MCS corresponds to Channel and Quality Indicator (CQI) value of 15 [1].

The remainder of this paper is organized as follows. In Section II, we present the system and channel model used in the simulation. In Section III, we present the MIMO schemes as defined in LTE. A brief review of the diversity schemes used in LTE systems is given in this section. A brief description of the Open Loop Spatial Multiplexing (OLSM) scheme is also reviewed in this section. The simulation results and discussion

 ISBN: 978-1-61208-203-5

of results will be presented in Section IV. Finally, we conclude our paper in Section V.

II. SYSTEM AND CHANNEL MODEL

In this section, the structure of the OFDM LTE signal is described. The OFDM signal has a time and a frequency domains. In the time domain, the LTE signal is composed of succesive frames. Each frame has a duration of 10 ms (T_{frame}). Each frame is divided into ten equally 1 ms long subframes. Each subrames consists of two equally long slots with 0.5 ms time duration (T_{slot}). For normal cyclic prefix length each slot consists of N_s = 7 OFDM symbols. In the frequency domain, the OFDM technique converts the LTE wide band signal into a number of narrowband signals. Each narrowband signal is transmitted on one subcarrier frequency. In LTE the spacing between subcarriers is fixed to 15 KHz. Tweleve adjacent subcarriers, occupying a total of 180 KHz, of one slot forms the so-called Resource Block (RB). The number of Resource Blocks in an LTE slot depends on the allowed system bandwidth. The minimum number of RB is equal to 6 corresponding to 1.4 MHz system bandwidth. For 20 MHz system bandwidth (Maximum Allowed bandwith in LTE) the number of RB is equal to 100. In MIMO system with M_R receive antenna and M_T transmit antenna, the relation between the received and the transmitted signals on subcarrier frequency k ($k \in 1, ...K$), at sampling instant time n ($n \in 1, ...N$) is given by

$$\mathbf{y}_{k,n} = \mathbf{H}_{k,n}\mathbf{x}_{k,n} + \mathbf{n}_{k,n} \tag{1}$$

$\mathbf{y}_{k,n} \in C^{M_R \mathrm{X}\ 1}$ is the received vector, $\mathbf{H}_{k,n} \in C^{M_R \mathrm{X}\ M_T}$ represents the channel matrix on subcarrier k at instant time n, $\mathbf{x}_{k,n} \in C^{M_T \mathrm{X}\ 1}$ is the transmit symbol vector and $\mathbf{n}_{k,n} \sim \mathcal{CN}(0, \sigma_n^2.\mathbf{I})$ is white, complex valued Gaussian noise vector with variance σ_n^2. Assuming perfect channel estimation, the channel matrix and noise variance are considered to be known at the receiver. A linear equalizer filter given by a matrix $\mathbf{F}_{k,n} \in C^{M_T \mathrm{X}\ M_R}$is applied on the received symbol vector $\mathbf{y}_{k,n}$ to determine the post-equalization symbol vector $\mathbf{r}_{k,n}$ [6]

$$\mathbf{r}_{k,n} = \mathbf{F}_{k,n}\mathbf{y}_{k,n} = \mathbf{F}_{k,n}\mathbf{H}_{k,n}\mathbf{x}_{k,n} + \mathbf{F}_{k,n}\mathbf{n}_{k,n} \tag{2}$$

The Zero Forcing (ZF) or Minimum Mean Square Error (MMSE) design criterion [7] are typically used for the linear receiver and the input signal vector is normalized to unit power. In MIMO-OFDM systems, the key factor of link error prediction and performances is the signal to noise ratio (SNR) which represents the measurement for the channel quality information. In practice, there are different measures and calculation procedures for the SNR in SISO and MIMO systems. In this study, the SNR is defined as follows [1]:

$$\gamma_{k,n} = \frac{\|\mathbf{H}_{k,n}\mathbf{x}_{k,n}\|_{\mathrm{F}}^2}{N_R\sigma_n^2} = \frac{N_R}{N_R\sigma_n^2} = \frac{1}{\sigma_n^2} \tag{3}$$

III. MIMO SCHEMES IN LTE

From theory it is well known that in MIMO systems the multiple antennas at the transmitter and the receiver can be used in two different modes, namely the diversity and multiplexing modes. Diversity mode can be used in the receiver (Receive Diversity) or at the transmitter (Transmit Diversity). Where receive diversity is simply a combining operation of different replica of the same transmitted signal, transmit diversity requires a space time coding operation of the transmitted signal. In LTE the two different modes are defined. In this section the different MIMO schemes defined in LTE are described.

A. Diversity Schemes

The transmit diversity techniques are defined only for 2 and 4 transmit antennas and one data stream. When two eNodeB antennas are available for transmit diversity operation, the Space Frequency Block Code (SFBC) [8] is used. SFBC is based on the well known Space Time Block Codes (STBC), also known as Alamouti codes [9]. STBC is defined in the UMTS and it operates on pairs of adjacent symbols in the time domain. As the signal in LTE is two dimensional (time and frequency domains) and the number of available OFDM symbols in a subframe is not always even, the direct application of STBC is not straightforward. In LTE for SFBC transmission, the symbols are transmitted from two eNodeB antenna ports on each pair of adjacent subcarriers as follows [8]:

$$\begin{bmatrix} y^{(0)}(1) & y^{(0)}(2) \\ y^{(1)}(1) & y^{(1)}(2) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ -x_2^* & x_1^* \end{bmatrix} \tag{4}$$

where $y^{(p)}(k)$ denotes the symbols transmitted on the k^{th} subcarrier from antenna port p. One important characteristic of such codes is that the transmitted signal streams are orthogonal and a simple linear receiver is required for optimal performances. Unfortunately, there is no known orthogonal codes for antenna configurations beyond 2 x 2 and the SFBC has been modified in order to be applied to the case of 4 transmit antennas. The new modified scheme of SFBC is known as Frequency Switched Transmit Diversity (FSTD). The frequency space code for 4 antennas is as follows:

 ISBN: 978-1-61208-203-5

$$\begin{bmatrix} y^{(0)}(1) & y^{(0)}(2) & y^{(0)}(3) & y^{(0)}(4) \\ y^{(1)}(1) & y^{(1)}(2) & y^{(1)}(3) & y^{(1)}(4) \\ y^{(2)}(1) & y^{(2)}(2) & y^{(2)}(3) & y^{(2)}(4) \\ y^{(3)}(1) & y^{(3)}(2) & y^{(3)}(3) & y^{(3)}(4) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & 0 & 0 \\ 0 & 0 & x_3 & x_4 \\ -x_2^* & x_1^* & 0 & 0 \\ 0 & 0 & -x_4^* & x_3^* \end{bmatrix} \quad (5)$$

The benefits of diversity can be exploited in different manners. It can increase the reliability of the radio link and it is quantified by the so called *diversity gain*. As a consequence of the diversity gain the error rate decreases. The data rate can also be improved logarithmically with respect to the number of antennas as antenna diversity increases the SNR linearly [10].

$$C = B\log_2(1+\gamma) \quad (6)$$

In addition, the coverage area can be improved or, for the same coverage area, the required power can be reduced. The *diversity gain* in MIMO systems is usually characterized by the number of independent fading diversity branches, also called *Diversity Order*. The diversity order is defined as the slope of the BLER versus SNR curve on a log-log scale. For a MIMO system with N_t transmit antennas and N_r receive antenna, it is said that the diversity order is N_d = N_t.N_r. The diversity order has a dramatic effect on the system reliability since the probability of one of the diversity branches having high SNR is higher compared to only one branche. In LTE, the SFBC (2x1) and FSTD (4x2) have a diversity order of 2 and 8 respectively.

B. Multiplexing Schemes

In Contrast to the diversity mode described in the previous section, the spatial multiplexing mode, which refers to splitting the incoming high data rate stream into N_t independent data streams, is considered, from a data throughput standpoint, as the most exciting type of MIMO systems. In MIMO system with N_t transmit antennas, the nominal spectral efficiency can be increased by a factor of N_t if the streams can be successfully and independently decoded. The factor N_t is known as *Multiplexing Gain*. In spatial multiplexing (N_t x N_r) MIMO system, the maximum data rate grows as [11]:

$$\min(N_t, N_r)\log(1+\gamma) \quad (7)$$

when γ is large.

In LTE, the spatial multiplexing mode is designated as Mode 3 and it is know as OLSM (Open Loop Spatial Multiplexing)

In SISO OFDM systems, the maximal data throughput depends on the available bandwidth and the parameter of the OFDM signal, like the number of subcarriers and the modulation order (QPSK, 16QAM, 64QAM). For a given frequency band (B) the maximal data throughput in bits per second can be approximated by the following simple equation [1]:

$$\textbf{Throughput}(bps) = \frac{N_{FB}.N_{SC}.N_{OFDM}.N_b.ECR}{T_{sub}} \quad (8)$$

where N_{FB} is the number of Frequency Block in the given frequency band (B); N_{SC} is the number of subcarrier in one Frequency Block; N_{OFDM} is the number of OFDM symbols in one subframe; N_b is the number of bits in one subcarrier; ECR is the Effective Code Rate, and T_{sub} is the duration of one subframe equal to 1 ms. In LTE, N_{SC} and N_{OFDM} are fixed and equals to 12 and 14 respectively [12]. For 5 MHz ($N_{FB} = 25$) bandwidth LTE system with 64 QAM Modulation ($N_b = 6$) and ECR = 0.9, the maximal data throughput that can be supported by the system is 22.68 Mbps.

B (MHz)	N_{FB}
1.4	6
5	25
10	50
15	75
20	100

IV. Simulation Results

In this section, we illustrate the results of the performances evaluation of three different MIMO schemes in 5 MHz LTE system using the MATLAB LTE link level simulator [1]. For comparison purpose, the performance of SISO scheme in the same system is also evaluated and presented. The three MIMO schemes are 2x1 (2 transmit antennas and only one receive antenna) SFBC diversity mode, 4x2 (4 transmit antennas and 2 receive antennas) FSTD diversity mode and 4x2 Open Loop Spatial Multiplexing (OLSM). The common simulation settings for the results are summarized in the next Table.

 ISBN: 978-1-61208-203-5

Parameter	Setting
Transmission Schemes	2x1 SFBC; 4x2 FSTD; 4x2 OLSM
Bandwidth	5 MHz
Simulation length	5000 subframes
Channel Type	Pedestrian B
Channel knowledge	perfect
CQI	15

The CQI value used in the simulation determines both the modulation order (64QAM) and the Effective Code Rate (0.92).

A. BLER Results

The Block Error Rate (BLER) results of SISO and MIMO schemes are shown in Figure 1. From the figure it is clear that the worst performances corresponds to the SISO curve (blue curve). The rate of change of the BLER in terms of SNR give us the estimation of the slope of the curve. As discussed in the previous sections, the slope of the BLER curve reflects the diversity order of the system. From the curve it can be observed that the slope is almost equal to one which means that the diversity order is equal to one as expected for the SISO configuration. As the modulation order is 64QAM a relatively high SNR is observed for the good BLER performance. An SNR of 41 dB is required to achieve a 10^{-3} value of BLER. The green curve represents the BLER results of the 2x1 diversity scheme. Asymptotically, the slope of this curve can be observed to be equal to two which corresponds to the diversity order of 2x1 system and hence a diversity gain of 2 as expected for 2x1 diversity scheme. An SNR gain can also be observed with respect to SISO scheme. In fact, it can be observed that to achieve a 10^{-2} value of BLER, the 2x1 diversity scheme needs about 8 dB less in SNR. In fact the BLER of 10^{-2} is achieved with 38 dB of SNR in SISO configuration however the same value of BLER is achieved with only 30 dB in the 2x1 diversity scheme. So an SNR gain of 8 dB is clearly observed for the 2x1 diversity scheme. The BLER results of the 4x2 Diversity scheme are represented by the red curve in Fig.1. In high SNR region the slope of the curve tends to be equal to 8. This value corresponds to the diversity order of a 4x2 system and hence a Diversity Gain of 8 can be observed from the curve. The SNR gain with respect to SISO configuration is more important than the case of 2x1 diversity scheme. In this case, an SNR gain of almost 18 dB at 10^{-2} value of BLER is obtained. Finally, the BLER results of the OLSM scheme are represented by the light blue curve and we can easily observe that the curve is almost parallel to the curve of 4x2 diversity scheme. This results is explained by the fact that the OLSM scheme uses the same antenna configuration as in the 4x2 diversity scheme and should have the same diversity order, which is equal to 8 (4x2) in this case. However the SNR gain is not the same as in 4x2 diversity mode but it is almost equal to SNR gain of 2x1 diversity system at 10^{-2} value of BLER (8 dB). This result is explained by the fact that in 4x2 OLSM scheme two different stream are sent from different antennas.

B. Data Throughput Results

The data throughput results of the three MIMO schemes are presented in the Fig.2 where they are compared to data throughput of SISO configuration. The data throughput of SISO configuration is shown by the blue curve. It can be observed that as the SNR increase the data throughput increase and it reaches its maximum at almost 40 dB. As in BLER results, the high order modulation is behind the high SNR required to achieve the maximum capacity. Beyond this value, the data throughput is constant and it corresponds to the maximum value as calculated in Section III-B. The green curve in Fig.2 represents the data throughput of the 2x1 diversity scheme. As expected there is no improvement in the data throughput as in 2x1 diversity scheme the same data is transmitted from the two antennas and no multiplexing gain can be achieved. However, the improvement comes from the fact that to achieve 15 Mbps, the 2x1 diversity scheme requires 5 dB less in SNR with respect to SISO configuration. In other words, the 15 Mbps is achieved by 30 dB SNR in SISO configuration and by only 25 dB in 2x1 diversity scheme. For the 4x2 diversity scheme, red curve in Fig.2, the improvement is even more and the gain in SNR is almost about 11 dB. It means that the 15 Mbps data throughput is reached by only 19 dB instead of 30 dB in SISO configuration. In this scheme also no multiplexing gain is observed as expected because as in the case of 2x1 diversity scheme only one signal stream is transmitted over the 4 transmit antennas. The multiplexing gain can easily be observed in the case of OLSM scheme, light blue curve. As in this scheme two different signal stream are transmitted simultaneously multiplexing gain of 2 is observed and the data throughput is almost doubled in high SNR.

V. Conclusion

In this paper, the performance evaluation of three different MIMO scheme in 5 MHz bandwidth LTE simulation using the MATLAB LTE simulator is presented. The improvement of these scheme with respect to SISO configuration is discussed. The difference between diversity mode and multiplexing mode and their respective gain in LTE MIMO schemes are also presented. The results clearly show an important improvement in terms of BLER and data throughput can be achieved in the three schemes.

References

[1] C. Mehlführer, M. Wrulich, J. C. Ikuno, D. Bosanska, and M. Rupp, "Simulating the long term evolution physical layer," in *Proc. of the 17th European Signal Processing Conference (EUSIPCO 2009)*, Glasgow, Scotland, Aug. 2009.

 ISBN: 978-1-61208-203-5

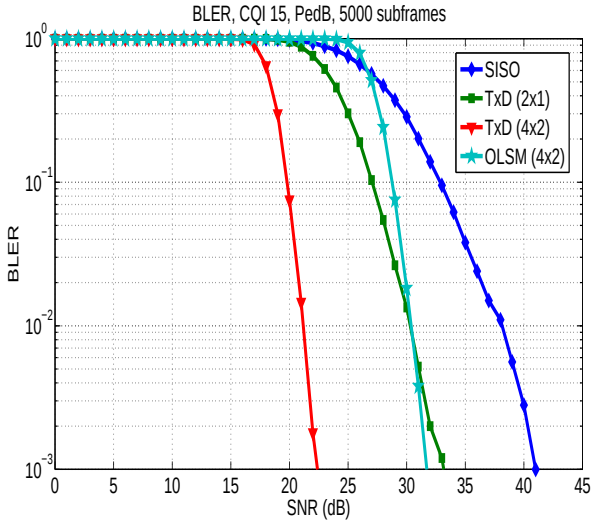


Fig. 1. BLER Performances of SISO and MIMO LTE Schemes

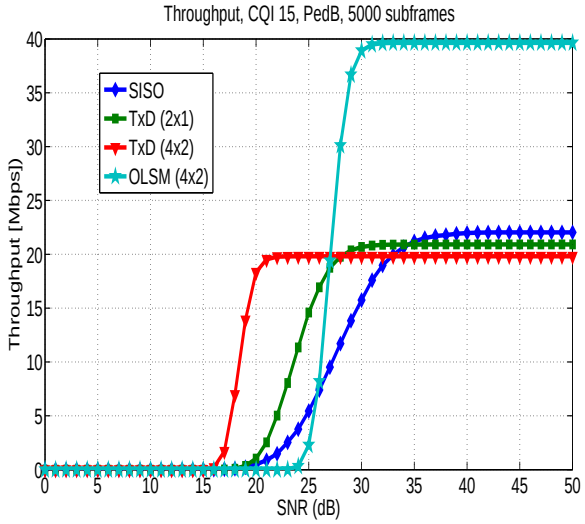


Fig. 2. Data Throughput of SISO and MIMO LTE schemes with 5 MHz bandwidth

[2] Online, available http://www.nt.tuwien.ac.at/ltesimulator.
[3] J. Ikuno, M. Wrulich, and M. Rupp, "System level simulation of lte networks," in *Vehicular Technology Conference (VTC 2010-Spring), 2010 IEEE 71st*, may 2010, pp. 1 –5.
[4] M. Simko, C. Mehlfuhrer, M. Wrulich, and M. Rupp, "Doubly dispersive channel estimation with scalable complexity," in *Smart Antennas (WSA), 2010 International ITG Workshop on*, feb. 2010, pp. 251 –256.
[5] S. Schwarz, M. Wrulich, and M. Rupp, "Mutual information based calculation of the precoding matrix indicator for 3gpp umts/lte," in *Smart Antennas (WSA), 2010 International ITG Workshop on*, feb. 2010, pp. 52 –58.
[6] S. Schwarz, C. Mehlfuhrer, and M. Rupp, "Calculation of the spatial preprocessing and link adaption feedback for 3gpp umts/lte," in *Wireless Advanced (WiAD), 2010 6th Conference on*, june 2010, pp. 1 –6.
[7] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2008.
[8] S. Sesia, T. Issam, and M. Backer, *LTE The UMTS Long Term Evolution From Theory To Practice*. John Wiley, 2011.
[9] S. Alamouti, "A simple transmit diversity technique for wireless communications," *Selected Areas in Communications, IEEE Journal on*, vol. 16, no. 8, pp. 1451 –1458, oct 1998.
[10] J. G. Andrews, A. Ghosh, and R. Muhamed, *Fundamentals of WiMAX: Understanding Broadband Wireless Networking*. Prentive HALL, 2007.
[11] N. Chiurtu, B. Rimoldi, and E. Telatar, "On the capacity of multi-antenna gaussian channels," in *Information Theory, 2001. Proceedings. 2001 IEEE International Symposium on*, 2001, p. 53.
[12] 3GPP, "Technical specification group radio access network," http://www.3gpp.org.

 ISBN: 978-1-61208-203-5

Joint Source-Relay Precoding with MMSE-based Interference Suppression in two-way MIMO Amplify and Forward Relays

Sundar Aditya
Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90089, USA
Email: sundarad@usc.edu

Rajeshwari S. S, K. Giridhar
Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai - 600036, India
Email: {rajeshwariss, giri}@tenet.res.in

***Abstract*—In this paper, we consider a two-way multiple input multiple output (MIMO) Amplify-and-Forward (AF) relay system, where interference is observed at the relay node during the multiple access (MAC) phase. For such a scenario, two new joint source-relay precoding algorithms with interference suppression are proposed and their performance analyzed through simulation results. These linear minimum mean-squared error (MMSE) based receiver algorithms provide acceptable error rate performance even in the presence of strong interference. Additionally, the effect of number of relay antennas and the number of interference streams on the overall diversity of the system is also investigated. We show that it is possible to handle interference at the relay node at the cost of losing some of the diversity gain offered by the extra antennas available at the relay.**

***Index Terms*—Two-way MIMO relay, Amplify-and-Forward, MMSE Interference Suppression, Joint Source-Relay Precoding**

I. Introduction

The use of relays in future wireless networks as a means to extend coverage as well as to improve the overall spectral efficiency has been gaining considerable attention recently. The principle of two-way relaying makes the use of relays spectrally efficient in spite of the additional hop, thus making it suitable for mass deployment in future wireless networks, where improving overall spectral efficiency is one of the goals.

While it is as yet not clear how relays will fit into the overall scheme of things in upcoming wireless networks, what can be inferred to a certain degree of confidence is that relays will mostly be operational in interference limited environments. One such plausible interference limited scenario is when a base-station (BS) acts as a relay between a user-pair and also receives information on the uplink from some other user, as illustrated in Fig. 1. To the user pair A, B (also referred to as sources) using the BS as a relay, the signal received by the BS from a different user C appears as interference. This is an example of interference affecting the performance of a relay during the MAC phase. A situation where interference is observed during the broadcast phase is shown in Fig. 2, where the user B sees interference from a neighboring BS serving its user D. It is important to observe from Fig. 1 and Fig. 2 that the cause of interference in the MAC and broadcast phases are different. Hence, MAC phase interference has an entirely different characteristics when compared to broadcast phase interference. Thus, the two kinds of interference are independent and therefore the two problems can be studied separately.

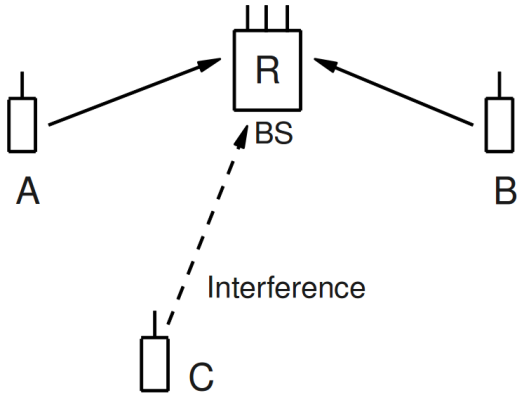


Fig. 1. Interference during MAC phase

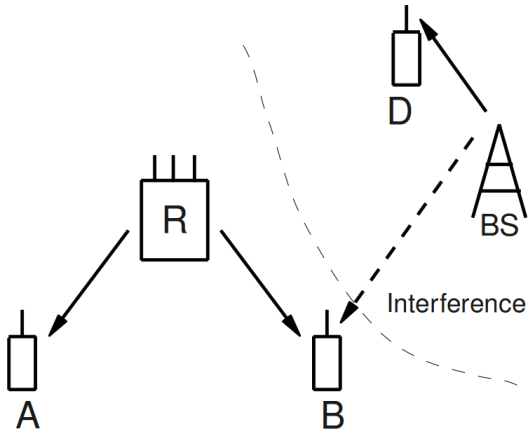


Fig. 2. Interference during broadcast phase

 ISBN: 978-1-61208-203-5

In this paper, we focus on the case where the performance of a two-way AF relay performance is hampered by interference during the MAC phase (Fig. 1). For such a situation, we are interested in devising an effective method to suppress the detrimental impact of interference. Furthermore, if all nodes have multiple antennas, additional gains can be achieved by means of MIMO precoding at both source and relay. Thus, in this work, we focus on the problem of joint source-relay precoding with interference suppression.

A. Prior Art

The joint design of source and relay precoders, along with source decoders is considered in [1] for a noise-limited system where only one stream of data is sent between the two communicating nodes utilizing the multi-antenna relay. However, with multi-antenna source and relay nodes, we would also like to transmit more than one stream of data from the sources. The problem of transmitting multiple streams has been studied in [2], where the source and relay precoders and decoders have been jointly optimized according to the AMSE (Arithmetic sum of Mean-Squared Error) as well as the ABER (Arithmetic sum of Bit-Error Rate) criteria for a purely noise-limited system without considering co-channel interference during the MAC phase. To the best of our knowledge, interference management in two-way MIMO relays has not been well-explored. Interference in MIMO relays has been handled in [3] in the limited context of separating the signal streams of multiple user pairs which are simultaneously using the same relay. In this work, we use a more generic model of interference and make an attempt to extend the framework proposed in [2] to cover the case where interference is also present in the system by. The key conclusion of this paper is that the joint source-relay precoding scheme designed for noise-limited systems in [2] can be used even in the presence of co-channel interference during the MAC phase as long as the relay performs MMSE-based interference suppression, for which only the knowledge of second-order statistics of interference is required.

B. Notation

Throughout this work, bold upper-case letters denote matrices (e.g., $\mathbf{X}, \mathbf{Y}$) while bold lower-case letters denote vectors (e.g., $\mathbf{x}$, $\mathbf{y}$). $(.)^T$, $(.)^H$ and $(.)^{-1}$ denote the transpose, hermitian (complex conjugate-transpose) and inverse of a matrix while $Tr\{.\}$ and $\mathbb{E}\{.\}$ represent the trace of a matrix and the expectation operator, respectively. CWGN stands for Circular White Gaussian Noise and $\mathcal{CN} \sim (\mu, K)$ represents a complex Gaussian random vector with mean μ and covariance matrix K.

This paper is divided into five sections. Section II details the signal model used, following which the joint source-relay MIMO precoding framework is presented in Section III. We back our hypothesis with simulation results presented in Section IV, and finally, Section V concludes the paper.

II. Signal Model

We interest ourselves in the problem where two multi-antenna transceiver nodes, A and B communicate with each other via an intermediate relay R. Let M_A, M_B and M_R denote the number of antennas at A, B and R, respectively. Such a configuration shall henceforth be referred to as a $M_A - M_R - M_B$ system. Let $\mathbf{H_A} \in \mathbb{C}^{M_R \times M_A}$ and $\mathbf{H_B} \in \mathbb{C}^{M_R \times M_B}$ represent the channels $A \longrightarrow R$ and $B \longrightarrow R$, respectively. Channel reciprocity is assumed to hold, whereby the channels $R \longrightarrow A$ and $R \longrightarrow B$ are represented by $\mathbf{H_A}^T$ and $\mathbf{H_B}^T$, respectively. A and B transmit L_A and L_B streams of data, respectively, where $L_A \leq \min(M_R, \min(M_A, M_B))$ and $L_B \leq \min(M_R, \min(M_B, M_A))$. The vectors $\mathbf{s_A} \in \mathbb{C}^{L_A \times 1}$ and $\mathbf{s_B} \in \mathbb{C}^{L_B \times 1}$ denote the symbols transmitted by A and B, respectively, and are assumed to contain independent, unit-energy symbols. The MIMO nature of the links can be exploited by performing precoding at A and B. Let $\mathbf{F_A} \in \mathbb{C}^{M_A \times L_A}$ and $\mathbf{F_B} \in \mathbb{C}^{M_B \times L_B}$ be the precoders used by A and B, respectively. Therefore, the signal at R seen after the first phase (MAC phase) is given by

$$\mathbf{y_R} = \mathbf{H_A}\mathbf{F_A}\mathbf{s_A} + \mathbf{H_B}\mathbf{F_B}\mathbf{s_B} + \mathbf{z} + \eta_R \tag{1}$$

where $\mathbf{z}$ denotes the interference seen at the relay and $\eta_R \sim \mathcal{CN}(0, \sigma_R^2 \mathbf{I}_{M_R})$ is the CWGN at the relay. We model the interference as streams of data transmitted by nodes other than A or B, i.e.,

$$\mathbf{z} = \sum_{i=1}^{k} \mathbf{H_i}\mathbf{s_i} \tag{2}$$

where we assume the presence of k streams of interference ($k \geq 1$), and $\mathbf{H_i} \in \mathbb{C}^{M_R \times 1}$ and $\mathbf{s_i} \in \mathbb{C}$ respectively denote the channel from the k^{th} interference stream to R, and the symbol transmitted by the k^{th} interference stream. Additionally, $\mathbf{F_A}$ and $\mathbf{F_B}$ satisfy the following constraints

$$Tr\{\mathbf{F_A}\mathbf{F_A}^H\} \leq P_A \tag{3}$$

$$Tr\{\mathbf{F_B}\mathbf{F_B}^H\} \leq P_B \tag{4}$$

where P_A and P_B represent the maximum average transmit powers at A and B, respectively. The signal $\mathbf{y_R}$ in (1) is amplified at R using the relay amplification matrix $\mathbf{G}$.

$$\tilde{\mathbf{y}}_\mathbf{R} = \mathbf{G}\mathbf{y_R} \tag{5}$$

This signal $\tilde{\mathbf{y}}_\mathbf{R}$ satisfies the following power constraint

$$Tr(\mathbb{E}\{\tilde{\mathbf{y}}_\mathbf{R}\tilde{\mathbf{y}}_\mathbf{R}^H\}) \leq P_R, \tag{6}$$

where P_R is the maximum power available at R.

In the second phase (broadcast phase), the relay signal $\tilde{\mathbf{y}}_R$ is transmitted to both A and B. The signals received at A and

B are as follows

$$\begin{aligned}\mathbf{y_A} &= \mathbf{H_A}^T\tilde{\mathbf{y}}_\mathbf{R} + \eta_A \\ &= \underbrace{\mathbf{H_A}^T\mathbf{G}\mathbf{H_A}\mathbf{F_A}\mathbf{s_A}}+\mathbf{H_A}^T\mathbf{G}\mathbf{H_B}\mathbf{F_B}\mathbf{s_B} + \mathbf{H_A}^T\mathbf{G}\mathbf{z} \\ &\quad + \mathbf{H_A}^T\mathbf{G}\eta_R + \eta_A \end{aligned} \tag{7}$$

$$\begin{aligned}\mathbf{y_B} &= \mathbf{H_B}^T\tilde{\mathbf{y}}_\mathbf{R} + \eta_B \\ &= \mathbf{H_B}^T\mathbf{G}\mathbf{H_A}\mathbf{F_A}\mathbf{s_A} + \underbrace{\mathbf{H_B}^T\mathbf{G}\mathbf{H_B}\mathbf{F_B}\mathbf{s_B}}+\mathbf{H_B}^T\mathbf{G}\mathbf{z} \\ &\quad + \mathbf{H_B}^T\mathbf{G}\eta_R + \eta_B \end{aligned} \tag{8}$$

Here, $\eta_A \sim \mathcal{CN}(0, \sigma_A^2\mathbf{I}_{M_A})$ and $\eta_B \sim \mathcal{CN}(0, \sigma_B^2\mathbf{I}_{M_B})$ denote CWGN at A and B, respectively, while the highlighted terms in (7) and (8) represent the self-interference seen at A and B, respectively. If $\mathbf{G}$ and $\mathbf{H_A}$ are known at A, then the self-interference seen by A can be subtracted, and similarly an equivalent condition holds at B too. We assume A, B and R to have perfect CSI of $\mathbf{H_A}$ and $\mathbf{H_B}$. Under such an assumption, it shall be seen that $\mathbf{G}$ can be computed in a decentralized manner at all the 3 nodes. The signals of interest therefore at A and B, after cancelling the back-propagating self-interference are as follows

$$\tilde{\mathbf{y}}_\mathbf{A} = \mathbf{H_A}^T\mathbf{G}\mathbf{H_B}\mathbf{F_B}\mathbf{s_B} + \mathbf{H_A}^T\mathbf{G}\mathbf{z} + \mathbf{H_A}^T\mathbf{G}\eta_R + \eta_A \tag{9}$$

$$\tilde{\mathbf{y}}_\mathbf{B} = \mathbf{H_B}^T\mathbf{G}\mathbf{H_A}\mathbf{F_A}\mathbf{s_A} + \mathbf{H_B}^T\mathbf{G}\mathbf{z} + \mathbf{H_B}^T\mathbf{G}\eta_R + \eta_B \tag{10}$$

A and B employ linear receivers $\mathbf{D_A}$ and $\mathbf{D_B}$ to get their respective estimates of $\mathbf{s_B}$ and $\mathbf{s_A}$.

$$\hat{\mathbf{s}}_\mathbf{B} = \mathbf{D_A}\tilde{\mathbf{y}}_\mathbf{A} \tag{11}$$

$$\hat{\mathbf{s}}_\mathbf{A} = \mathbf{D_B}\tilde{\mathbf{y}}_\mathbf{B} \tag{12}$$

III. Joint Source-Relay MIMO Precoding

From (9) and (11), the expression for the MSE matrix at A is given by

$$\begin{aligned}MSE_A &= \mathbb{E}\{(\mathbf{s_B} - \hat{\mathbf{s}}_\mathbf{B})(\mathbf{s_B} - \hat{\mathbf{s}}_\mathbf{B})^H\} \\ &= \mathbf{I} + \mathbf{D_A}\mathbb{E}\{\tilde{\mathbf{y}}_\mathbf{A}\tilde{\mathbf{y}}_\mathbf{A}^H\}\mathbf{D_A}^H - \mathbf{D_A}\mathbf{H_A}^T\mathbf{G}\mathbf{H_B}\mathbf{F_B} \\ &\quad - (\mathbf{H_A}^T\mathbf{G}\mathbf{H_B}\mathbf{F_B})^H\mathbf{D_A}^H \end{aligned} \tag{13}$$

where $\tilde{\mathbf{y}}_\mathbf{A}$ is as given in (9). For the rest of this section, we consider node A while presenting our analysis. The results for B can be obtained quite easily by the symmetry of the problem. For fixed $\mathbf{G}$, $\mathbf{F_A}$ and $\mathbf{F_B}$, the optimal linear receiver $\mathbf{D_A}$ in terms of minimizing the MSE can be obtained by evaluating $\nabla_{\mathbf{D_A}}(MSE_A) = 0$, which yields the familiar Wiener-filter solution.

$$\begin{aligned}\mathbf{D_A} &= (\mathbf{H_A}^T\mathbf{G}\mathbf{H_B}\mathbf{F_B})^H\mathbb{E}\{\tilde{\mathbf{y}}_\mathbf{A}\tilde{\mathbf{y}}_\mathbf{A}^H\}^{-1} \\ &= (\mathbf{H_A}^T\mathbf{G}\mathbf{H_B}\mathbf{F_B})^H[(\mathbf{H_A}^T\mathbf{G}\mathbf{H_B}\mathbf{F_B})(\mathbf{H_A}^T\mathbf{G}\mathbf{H_B}\mathbf{F_B})^H \\ &\quad + (\mathbf{H_A}^T\mathbf{G})\mathbf{R_{i+n}}(\mathbf{H_A}^T\mathbf{G})^H + \sigma_a^2\mathbf{I}_{M_A}]^{-1} \end{aligned} \tag{14}$$

where $\mathbf{R_{i+n}} = \mathbb{E}\{\mathbf{z}\mathbf{z}^H\} + \sigma_R^2\mathbf{I}_{M_R}$ is the covariance matrix of the interference plus noise, at the relay. It is demonstrated in [4] that the BER is a convex increasing function of the MSE for small values of the argument (for BER less than 2×10^{-2} as a thumb rule). Thus, we are justified in our choice of a linear MMSE receiver as it is not only easy to implement but also ensures good BER performance for most practical cases.

Substituting (14) in (13) and using the matrix inversion lemma, the following expression is obtained for the MSE matrix at A

$$MSE_A = (\mathbf{I} + \mathbf{F_B}^H\mathbf{R_B}\mathbf{F_B})^{-1} \tag{15}$$

where $\mathbf{R_B} = (\mathbf{H_A}^T\mathbf{G}\mathbf{H_B})^H[(\mathbf{H_A}^T\mathbf{G})\mathbf{R_{i+n}}(\mathbf{H_A}^T\mathbf{G})^H + \sigma_A^2\mathbf{I}_{M_A}]^{-1}(\mathbf{H_A}^T\mathbf{G}\mathbf{H_B})$ We now focus our attention to the design of $\mathbf{F_A}$, $\mathbf{F_B}$ and $\mathbf{G}$ according to the AMSE and ABER optimization criteria, as specified in [2].

The AMSE and ABER objective functions have the following form:

(i) AMSE

$$f_{AMSE} = Tr\{MSE_A\} + Tr\{MSE_B\} \tag{16}$$

(ii) ABER

$$\begin{aligned}f_{ABER} &= \sum_{i=1}^{L_B} BER_{A_i} + \sum_{j=1}^{L_A} BER_{B_j} \\ &= \sum_{i=1}^{L_B} Q(\sqrt{2(MSE_{A_{i,i}}^{-1} - 1)}) + \\ &\quad \sum_{i=1}^{L_A} Q(\sqrt{2(MSE_{B_{i,i}}^{-1} - 1)}\ ) \end{aligned} \tag{17}$$

where BER_{A_i} and BER_{B_j} respectively denote the BERs for the i^{th} stream at A and the j^{th} stream at B, and $Q(.)$ denotes the Q-function with (17) being valid for QPSK constellation [5] at A and B, and the summation is over the number of streams transmitted by A and B.

We proceed in an iterative manner to converge upon the source precoders as well as $\mathbf{G}$. Firstly, for fixed $\mathbf{G}$, the optimization problems that need to be solved for the AMSE and ABER criteria are as follows

(i) AMSE criterion

$$\begin{aligned}&\min_{\mathbf{F_i}|i=A,B}\ f_{AMSE} \\ &\text{subject to} \\ &Tr\{\mathbf{F_i}\mathbf{F_i}^H\} \le P_i \end{aligned} \tag{18}$$

(ii) ABER criterion

$$\begin{aligned}&\min_{\mathbf{F_i}|i=A,B}\ f_{ABER} \\ &\text{subject to} \\ &Tr\{\mathbf{F_i}\mathbf{F_i}^H\} \le P_i \end{aligned} \tag{19}$$

For fixed $\mathbf{G}$, the design of source precoders $\mathbf{F_A}$ and $\mathbf{F_B}$ becomes decoupled. We present the solution for $\mathbf{F_B}$. The solution for $\mathbf{F_A}$ can be obtained in an identical manner. The optimal precoder structures for (18) and (19), as demonstrated in [2], is given by

 ISBN: 978-1-61208-203-5

(i) AMSE

$$\mathbf{F_B} = \mathbf{U_B}\mathbf{\Sigma_B} \tag{20}$$

(ii) ABER

$$\mathbf{F_B} = \mathbf{U_B}\mathbf{\Sigma_B}\mathbf{V}^H \tag{21}$$

Here, $\mathbf{U_B} \in \mathbb{C}^{M_B \times L_B}$ contains the left eigenvectors corresponding to the highest L_B eigenvalues of $\mathbf{R_B}$ in ascending order, and $\mathbf{\Sigma_B} \in \mathbb{C}^{L_B \times L_B}$ denotes the diagonal matrix containing the corresponding powers allocated to the various streams. The water-filling algorithm used to allocate the powers is given in [4] and $\mathbf{V} \in \mathbb{C}^{L_B \times L_B}$ is any unitary matrix like the DFT matrix or the Hadamard matrix.

Given $\mathbf{F_A}$ and $\mathbf{F_B}$, f_{AMSE} and f_{ABER} become non-linear functions of $\mathbf{G}$, with quadratic constraints involving $\mathbf{G}$. A closed form solution for $\mathbf{G}$ is as yet unknown. Thus, we resort to numerical techniques and propose two SQP (sequential quadratic programming) based methods for the design of the relay precoder - i) with implicit interference suppression and ii) with explicit interference suppression.

A. Relay Precoder with Implicit Interference Suppression

In the analysis presented so far, the effects of interference suppression as well as relay precoding are combined into one effective relay amplification matrix $\mathbf{G}$. Thus, the optimization problems for the AMSE and the ABER criteria have the following form

(i) AMSE criterion

$$\begin{aligned} &\min_{\mathbf{G}} \ f_{AMSE} \\ &\text{subject to} \\ &Tr\{\mathbf{G}[\mathbf{H_A}\mathbf{F_A}\mathbf{F_A}^H\mathbf{H_A}^H + \mathbf{H_B}\mathbf{F_B}\mathbf{F_B}^H\mathbf{H_B}^H \\ &\quad + \mathbf{R_{i+n}}]\mathbf{G}^H\} \leq P_R \end{aligned} \tag{22}$$

(ii) ABER criterion

$$\begin{aligned} &\min_{\mathbf{G}} \ f_{ABER} \\ &\text{subject to} \\ &Tr\{\mathbf{G}[\mathbf{H_A}\mathbf{F_A}\mathbf{F_A}^H\mathbf{H_A}^H + \mathbf{H_B}\mathbf{F_B}\mathbf{F_B}^H\mathbf{H_B}^H \\ &\quad + \mathbf{R_{i+n}}]\mathbf{G}^H\} \leq P_R \end{aligned} \tag{23}$$

For (22) and (23), the solution for $\mathbf{G}$ is obtained through SQP. The joint source-relay precoding algorithm with implicit interference suppression is summarized in Table I.

B. Relay Precoder with Explicit Interference Suppression

Since a closed form solution for the relay precoder $\mathbf{G}$ is unavailable, the interference suppressing capabilities of the relay precoders obtained as solutions to (22) and (23) may be restricted. In this section, we propose the use of MMSE-based interference suppression at the relay to explicitly take care of the interference, before amplifying the desired signal components using the relay precoder $\mathbf{G}$. The signals of interest therefore at A and B at the end of the second phase are as follows:

Step 1	Set $k = 1$. Fix $\mathbf{G}_k = \gamma_R \mathbf{I}_{M_R}$, $\mathbf{F_A}_k = \gamma_A \mathbf{I}_{M_A}$ and $\mathbf{F_B}_k = \gamma_B \mathbf{I}_{M_B}$, where $\gamma_A = \sqrt{P_A/M_A}$, $\gamma_B = \sqrt{P_B/M_B}$ and $\gamma_R = \sqrt{P_R/Tr\{\mathbf{H_A}\mathbf{F_A}\mathbf{F_A}^H\mathbf{H_A}^H + \mathbf{H_B}\mathbf{F_B}\mathbf{F_B}^H\mathbf{H_B}^H + \mathbf{R_{i+n}}\}}$ (uniform power allocation).
Step 2	Compute $\mathbf{F_B}_{k+1}$ and $\mathbf{F_A}_{k+1}$ using $\mathbf{G}_k$, $\mathbf{H_A}$ and $\mathbf{H_B}$ according to (20) and (21) for the AMSE and ABER criterion, respectively.
Step 3	Using $\mathbf{F_A}_{k+1}$ and $\mathbf{F_B}_{k+1}$, solve for $\mathbf{G}_{k+1}$ by SQP as shown in (22) and (23), for the AMSE and ABER criterion, respectively.
Step 4	If $\lvert f_{AMSE_{k+1}} - f_{AMSE_k}\rvert \geq \epsilon$ for the AMSE criterion, or if $\lvert f_{ABER_{k+1}} - f_{ABER_k}\rvert \geq \epsilon$ for the ABER criterion, then set $k = k + 1$ and repeat from step 2 onwards.

TABLE I
SUMMARY OF JOINT SOURCE-RELAY PRECODING WITH IMPLICIT INTERFERENCE SUPPRESSION

$$\tilde{\mathbf{y}}_\mathbf{A} = \mathbf{H_A}^T\mathbf{G}\mathbf{W}\mathbf{H_B}\mathbf{F_B}\mathbf{s_B} + \mathbf{H_A}^T\mathbf{G}\mathbf{W}\mathbf{z} + \mathbf{H_A}^T\mathbf{G}\mathbf{W}\eta_R + \eta_A \tag{24}$$

$$\tilde{\mathbf{y}}_\mathbf{B} = \mathbf{H_B}^T\mathbf{G}\mathbf{W}\mathbf{H_A}\mathbf{F_A}\mathbf{s_A} + \mathbf{H_B}^T\mathbf{G}\mathbf{W}\mathbf{z} + \mathbf{H_B}^T\mathbf{G}\mathbf{W}\eta_R + \eta_B \tag{25}$$

where $\mathbf{W}$ denotes the MMSE-interference suppression matrix, acting on the signal at the relay at the end of the first phase before the relay precoder $\mathbf{G}$. $\mathbf{W}$ has the following form:

$$\mathbf{W} = (\mathbf{H_A}\mathbf{F_A}\mathbf{F_A}^H\mathbf{H_A}^H + \mathbf{H_B}\mathbf{F_B}\mathbf{F_B}^H\mathbf{H_B}^H)\mathbf{R_{i+n}}^{-1} \tag{26}$$

The solutions for $\mathbf{F_A}$, $\mathbf{F_B}$, $\mathbf{G}$ and $\mathbf{W}$ are jointly computed iteratively in a manner similar to the one described in the above section. For fixed $\mathbf{G}$ and $\mathbf{W}$, the source precoder $\mathbf{F_B}$ is computed as given in (20) and (21), except that $\mathbf{R_B}$ has the following structure instead of the one given in the previous section

$$\mathbf{R_B} = (\mathbf{H_A}^T\mathbf{G}\mathbf{W}\mathbf{H_B})^H[(\mathbf{H_A}^T\mathbf{G}\mathbf{W})\mathbf{R_{i+n}}(\mathbf{H_A}^T\mathbf{G}\mathbf{W})^H + \sigma_A^2\mathbf{I}_{M_A}]^{-1}(\mathbf{H_A}^T\mathbf{G}\mathbf{W}\mathbf{H_B})$$

The expression for $\mathbf{F_A}$ can be obtained in a similar manner as well. For fixed $\mathbf{F_A}$ and $\mathbf{F_B}$, $\mathbf{W}$ can be computed as given in (26), and $\mathbf{G}$ can then be computed by SQP using the knowledge of $\mathbf{F_A}$, $\mathbf{F_B}$ and $\mathbf{W}$, similar to (22) and (23). The joint source-relay precoding algorithm with explicit MMSE-based interference suppression at the relay is summarized in Table II.

It can be observed in both the above-mentioned methods that, with perfect knowledge of the channels $\mathbf{H_A}$ and $\mathbf{H_B}$, all three nodes can independently run the above mentioned algorithms at their end and arrive at the same set of source and relay precoders. Of course, the second-order statistics of the interference, i.e., $\mathbf{R_{i+n}}$, need to be made available at both A and B as well. However, since the covariance matrix is Hermitian Toeplitz, the amount of overhead required to communicate it is quite small and can be easily accomplished.

 ISBN: 978-1-61208-203-5

Step 1	Set $k = 1$. Fix $\mathbf{F_{A}}_k = \gamma_A \mathbf{I}_{M_A}$, $\mathbf{F_{B}}_k = \gamma_B \mathbf{I}_{M_B}$, where $\gamma_A = \sqrt{P_A/M_A}$, $\gamma_B = \sqrt{P_B/M_B}$. Set $\mathbf{W}_k = (\mathbf{H_A F_{A}}_k \mathbf{F_{A}}_k^H \mathbf{H_A}^H + \mathbf{H_B F_{B}}_k \mathbf{F_{B}}_k^H \mathbf{H_B}^H)\mathbf{R_{i+n}}^{-1}$. Fix $\mathbf{G_k} = \gamma_R \mathbf{I}_{M_R}$, where $\gamma_R = \sqrt{\frac{P_R}{Tr\{\mathbf{W}_k[\mathbf{H_A F_A F_A}^H \mathbf{H_A}^H + \mathbf{H_B F_B F_B}^H \mathbf{H_B}^H + \mathbf{R_{i+n}}]\mathbf{W}_k^H\}}}$.
Step 2	Compute $\mathbf{F_{B}}_{k+1}$ and $\mathbf{F_{A}}_{k+1}$ using $\mathbf{G}_k$, $\mathbf{W}_k$, $\mathbf{H_A}$ and $\mathbf{H_B}$ according to (20) and (21), for the AMSE and ABER criterion, respectively.
Step 3	Using $\mathbf{F_{A}}_{k+1}$ and $\mathbf{F_{B}}_{k+1}$, compute $\mathbf{W}_{k+1}$ as given in (26).
Step 4	Using $\mathbf{F_{A}}_{k+1}$, $\mathbf{F_{B}}_{k+1}$ and $\mathbf{W}_{k+1}$, solve for $\widehat{\mathbf{G}}_{k+1}$ by SQP as shown in (22) and (23), for the AMSE and ABER criterion, respectively.
Step 5	If $\lvert f_{AMSE_{k+1}} - f_{AMSE_k} \rvert \geq \epsilon$ for the AMSE criterion, or if $\lvert f_{ABER_{k+1}} - f_{ABER_k} \rvert \geq \epsilon$ for the ABER criterion, then set $k = k + 1$ and repeat from step 2 onwards.

TABLE II
SUMMARY OF JOINT SOURCE-RELAY PRECODING WITH EXPLICIT (MMSE) INTERFERENCE SUPPRESSION

IV. SIMULATION RESULTS

In this section, we compare the relative performance of implicit and explicit interference suppression at the relay using simulation results. The simulations have been carried out using MATLAB. For a $2-4-2$ case, the MSE performance of both methods is shown in Fig. 3 for AMSE-based precoding. Here, A and B, each having 2 antennas, transmit 2 streams of information each to the relay during the first phase ($L_A = L_B = 2$), along with the presence of a single interferer who is also transmitting 2 streams of information. Hence, the relay, with 4 antennas, receives 6 streams of information during Phase 1. For the same $2-4-2$ configuration, Fig. 4 contains the BER performance of both methods for ABER precoding. For the simulations, Rayleigh fading channels have been assumed for $\mathbf{H_A}$, $\mathbf{H_B}$ and $\mathbf{H_i}$ (all $\in \mathbb{C}^{4\times 2}$) with all channel coefficients being drawn from $\mathcal{CN} \sim (0, 1)$. A and B use QPSK constellation for their symbols $\mathbf{s_A}$ and $\mathbf{s_B}$ ($\in \mathbb{C}^{2\times 1}$) and equal power constraints are assumed at all 3 nodes ($P_A = P_B = P_R$). The interferer transmits symbols $\mathbf{s_i} \in \mathbb{C}^{(2x1)}$ having the following property: $\mathbb{E}\{\mathbf{s_i s_i}^H\} = P_{intf}\mathbf{I}$, where P_{intf} denotes the interferer's power and $\mathbf{I}$, the identity matrix. SIR has been defined as $10\log(P_A/P_{intf})$. Additionally, equal noise floors are also assumed at A, B and R ($\sigma_A^2 = \sigma_B^2 = \sigma_R^2 = \frac{P_{intf}}{1000}$), with the noise floor being 30dB below P_{intf} in order to make the system interference-limited. The simulation results have been obtained by averaging over 1000 independent channel realizations.

We observe from Fig. 3 and Fig. 4 that explicit MMSE-based interference suppression at the relay yields better BER and MSE performance when compared to implicit interference suppression using SQP based relay precoders. It can also be observed from the BER curves in Fig. 4 that a $2-4-2$ system is capable of sending 2 streams of information each from A and B in addition to suppressing 2 streams of interference.

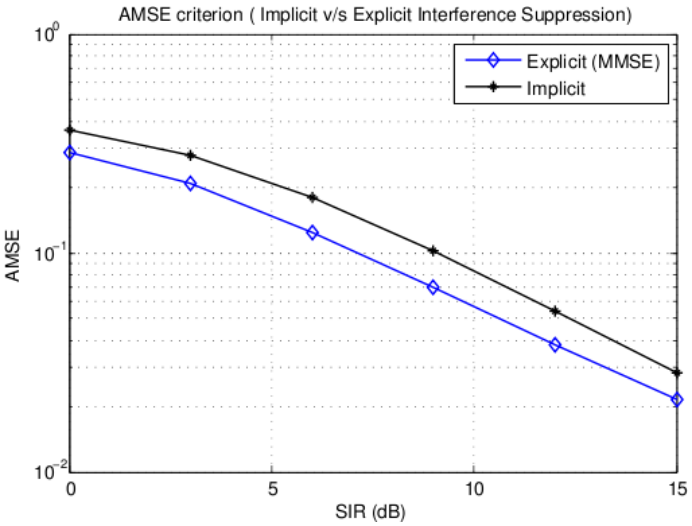


Fig. 3. MSE performance of Implicit and Explicit Interference Suppression at the relay

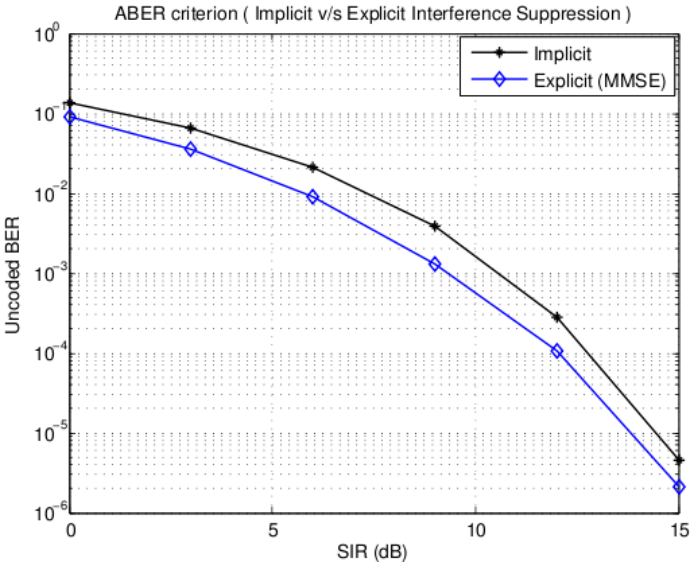


Fig. 4. BER performance of Implicit and Explicit Interference Suppression at the relay

While it is well known that a base station with 4 antennas can handle 4 streams of information in the uplink, it can be gauged from Fig. 4 that when acting as a relay between 2 user pairs, a base station with 4 antennas is capable of handling 6 streams of information on the uplink. Thus, it is of interest to study the effect of number of relay antennas on the diversity order that can be achieved as far as BER performance is concerned.

A. Impact of Relay Antennas on Diversity Order

With A and B transmitting 2 streams of information each, the BER performance of a $2-4-2$ system with 2 streams of interference is compared with that of a $2-5-2$ system with 3 streams of interference and a $2-6-2$ system with 4 streams of interference in Fig. 5. It can be observed from eye-balling the BER curves that they all have the same slope.

 ISBN: 978-1-61208-203-5

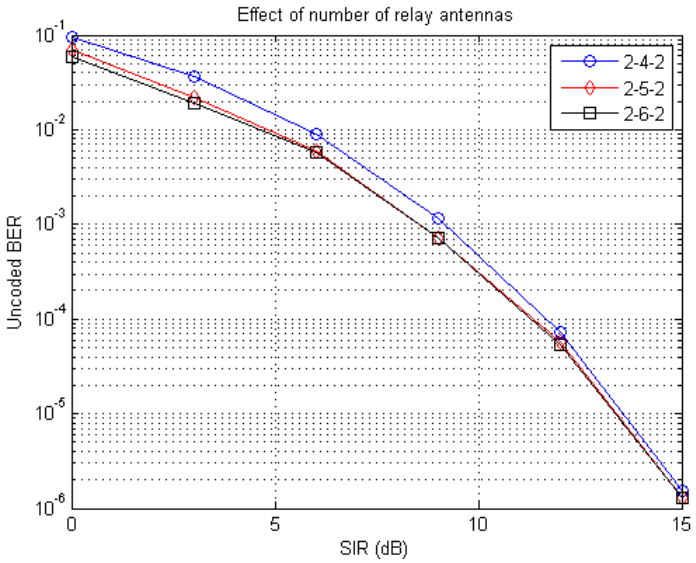


Fig. 5. Diversity-multiplexing trade-off at the relay

Hence, a $2-5-2$ system is able to suppress an extra stream of interference while achieving the same diversity order of a $2-4-2$ system. Likewise, a $2-6-2$ system achieves the same diversity order of a $2-4-2$ system while suppressing 2 additional streams of interference. Thus, in general, additional antennas at the relay help in mitigating more interference at the cost of a loss in diversity. This is a manifestation of the diversity-multiplexing tradeoff in a situation where co-channel interference hampers relay communication during the MAC phase.

V. Conclusion

In this paper, a scenario where relay operation is hampered by the presence of interference in the first MAC phase was considered. In such a situation, the effectiveness of MMSE-based interference suppression at the relay along with joint source-relay precoding was demonstrated using simulation results. It needs to be noted that while there is still some residual interference at the relay even after MMSE interference suppression, the fact that the joint source-relay precoding takes the residual interference into account is what makes our scheme robust even in low SINR regimes. Thus, the joint source-relay precoding scheme proposed in [2] for noise-limited systems is effective even in an interference-limited scenario provided the relay performs MMSE-based interference suppression, for which only the knowledge of second-order statistics of interference is required. The effect of number of relay antennas on the overall diversity order of BER performance in the presence of interference was also studied using simulation results for various configurations, where it was observed that with additional antennas at the relay, it is possible to suppress more streams of interference at the cost of loss in diversity.

References

[1] F. Roemer and M. Haardt, "Algebraic Norm-Maximizing (ANOMAX) Transmit Strategy for Two-Way Relaying With MIMO Amplify and Forward Relays," *IEEE Signal Processing Letters*, vol. 16, no. 10, pp. 909–912, Oct. 2009.

[2] S. Rajeshwari and K. Giridhar, "New Approach to Joint MIMO Precoding for 2-way AF Relay Systems," in *Proc. 17th National Conference on Communications*, Bangalore, India, Jan. 2011, pp. 1–5.

[3] A. U. T. Amah and A. Klein, "Pair-Aware Transceive Beamforming for Non-Regenrative Multi-User Two-Way Relaying," in *Proc. IEEE ICASSP*, Dallas, USA, Jan. 2010, pp. 2506–2509.

[4] D. P. Palomar, J. M. Cioffi, and M. A. Lagunas, "Joint Tx-Rx Beamforming Design for Multicarrier MIMO Channels: A Unified Framework for Convex Optimization," *IEEE Trans. Signal Process.*, vol. 51, no. 9, pp. 2381–2401, Sep. 2003.

[5] J. G. Proakis and M. Salehi, *Digital Communications*, 5th ed. McGraw-Hill Higher Education, 2008.

 ISBN: 978-1-61208-203-5

Monitoring of Environmental Parameters in Nanoelectronic Fabrication

Mokhloss I. Khadem, Valentin Sgarciu
Faculty of Automatic Control and Computer Science
"Politehnica" University of Bucharest
Bucharest, Romania
sml_ka@yahoo.com, vsgarciu@aii.pub.ro

***Abstract*—Nanoelectronics are the essential hardware enabler for electronic product and service innovation in key growth markets for global industry, such as telecommunications, transportation and medical technology. dust particles, humidity , temperature parameters have big effect for the fabrication area due to the newest technologies that are usually very complex and very sensitive to the external influences. Dust is one of the major problems of the top-down lithographic approaches is that they require very clean environments because dust and particulates can mask part of the exposed area. For high yields and hence low costs, clean rooms must have particulate densities extremely low and at sizes much smaller than the lithographic minimum feature size. In this context, this paper presents a data acquisition system that is capable to monitor and measure three types of environmental parameters: dust, temperature, and humidity. The remote board containing sensors and processing circuits is connected to the server through wireless network. The analogue to digital conversion is realized by data acquisition card (MDA300). The systems can operate for monitoring and control. The sensors communicate wirelessly with selectable sensors via three wireless smart sensors: temperature, dust sensor and humidity sensor. The software control and the acquired data processing are realized with a MOTEVIEW application that is capable to simultaneously display the measured data. This paper introduces a prototype for an affordable wireless sensors network for monitoring air quality. It can work in two modes: online and offline. In online mode, the sensors periodically send the readings to the base station. In offline mode, the sensors store the readings periodically to the internal memory and these readings can be collected whenever they are needed. With minimal changes, the proposed system can be extended to operate with more types of sensors. Using a wireless transmission method between PC and remote board, the operation distance of the system can be further extended. We obtained encouraging results regarding the accuracy of the optical measurements from dust sensors connected in the wireless network**

Keywords-dust sensors; temperature and humidity sensors; wireless network sensors; Mote-VIEW software

I. INTRODUCTION

In order to monitor and control dust, humidity and temperature in the fabrication processes of electronic components and modules has gained an increasing importance due to the complexity and sensitivity of the new technologies to the influences generated by ambient humidity, dust particles, temperature, radiation levels, pressure, etc.

Due to these facts, when we refer to IC (Integrated Circuits) fabrication technologies for example, where nanotechnologies are currently in use, the control of environmental conditions has become mandatory. Also, in the case of printed circuit boards (PCB) an exposure to humid ambient conditions for example will cause the absorption of moisture that can greatly affect the proper operation of the equipment that contains the respective module. The corrosion on metallic parts of an electronic assembly is another problem generated by the humidity, especially in the fabrication stage This paper presents a data acquisition system with Mote-VIEW software can show the result of measurement parameters and the chart view of the network to monitor the most of important parameters that characterize the environment for PCB fabrication: humidity, dust, temperature. The related work is a data acquisition system that is capable to monitor and measure some environmental parameters like: pressure, temperature and humidity, based on Labview software without using wireless sensor network .The presented system is a scalable industrial quality monitoring system for clean rooms and can be also used as a component of a more complex equipment intended for testing and reliability evaluation of electronic modules in different environmental conditions.

The environmental effects of dust in nanoelectronic fabrication will be described in the 2nd section of this paper, following in the 3rd section with the presentation of a wireless sensor system proposed for the measurement and monitor of the dust, humidity, temperature in a fabrication clean room. In the 4th section we describe the monitoring software designed in Mote-view for the hardware system, followed by the final chapter with the conclusions.

II. ENVIROMENT EFFECTS

Nanoelectronic fabrication technologies originate from the microelectronics industry, and the devices are usually made on silicon wafers even though glass, plastics and many other substrates are in use, microelectronics extension into nanoscale (for example NEMS, for nano electro mechanical systems) have re-used, adapted or extended micro fabrication methods. Micro fabrication is known as "semiconductor manufacturing" or "semiconductor device fabrication is actually a collection of technologies which are utilized in making microdevices [1]. Micro fabrication is carried out in clean rooms, where air has been filtered of particle contamination and temperature, humidity, vibrations and electrical disturbances are under stringent control.

Smoke, dust, bacteria and cells are micrometers in size, and their presence will destroy the functionality of a micro fabricated device. Wafer cleaning and surface preparation work a little bit like the machines in a bowling alley: first they

 ISBN: 978-1-61208-203-5

remove all unwanted bits and pieces, and then they reconstruct the desired pattern so that the game can go on [2].

Moisture can accelerate various failure mechanisms in printed circuit board assemblies. Moisture can be initially present in the epoxy glass prepare , absorbed during the wet processes in printed circuit board manufacturing, or diffuse into the printed circuit board during storage. Moisture can reside in the resin, resin/glass interfaces, and micro-cracks or voids due to defects [3].

Higher reflow temperatures associated with lead-free processing increase the vapor pressure, which can lead to higher amounts of moisture uptake compared to eutectic tin-lead reflow processes. The processing of silicon wafers to produce integrated circuits (IC) involves specific chemistry and physics to build up a succession of layers of materials and geometries to produce thousands of electronic devices at very small dimensions. The conditions under which these processes can work to successfully to transform the silicon into ICs require the absence of contaminants (dust, humidity, unwanted chemical elements etc.). Thus, the process chambers normally operate under vacuum, with elemental, molecular, and other particulate contaminants rigorously controlled. The ideas presented in the above sections support the necessity of measurement and monitoring systems for environmental conditions in fabrication areas [5].

III. SYSTEM ARCHITECTURE

To measure and monitor the dust, humidity, temperature values in a fabrication clean room, we propose the distribution of several nodes, from N_1 to N_n. Each node is a smart sensor operating in a Plug-and-Play mode and each node communicates to a server, over a wireless network by using the IEEE 1451.5-802.11 standard [4].

This standard will enable sensors and devices to communicate wirelessly, eliminating the monetary and time costs of installing cables at acquisition points. IEEE is currently working on three different standards, IEEE 802.15.4, Bluetooth and Zigbee.

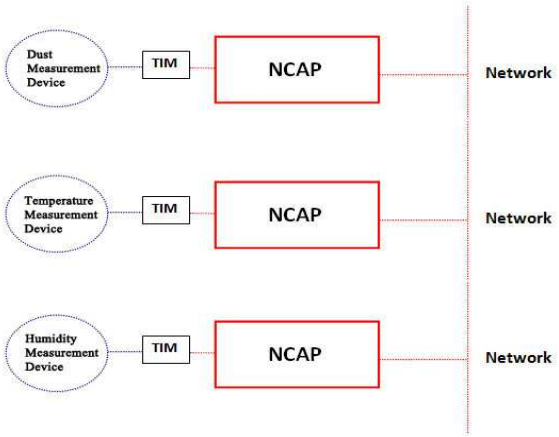


Figure 1. Dust, temperature, humidity smart sensors.

In the proposed implementation, 3 sensors are: Sharp GP2Y1010AU0F dust sensor that is based on the optical principle [6] and LM35 for temperature [11] and HIH 3605 for humidity level measurements [10] connected with MDA 300 data acquisition [12] from crossbow and these sensors communicate wirelessly to NI wireless sensor network (WSN) from National Instruments devices provide the same quality and accuracy as traditional wired measurement systems [14], but with increased flexibility, lower costs, and the ability to create smart WSN systems based on Mote-VIEW software [7].

Mote-VIEW is designed to be an interface ("client layer") between a user and a deployed network of wireless sensors. *Mote-VIEW* provides users the tools to simplify deployment and monitoring. It also makes it easy to connect to a database, to analyze, and to graph sensor. The humidity sensor (HIH 3605) consists of a polymer capacitive sensing element with on-chip integrated signal conditioning and a second polymer layer to protect against dirt. The humidity sensor has a linear voltage output with an accuracy of ±2% RH (relative humidity) and ±0.5% RH linearity. If the measurement is realized in slowly moving air at 25°C the response time of this sensor is of maximum 15s. The LM35 is calibrated directly in Celsius degrees and has an sensitivity of a 10.0 mV/°C and an 0.5°C accuracy over -55°C to +150°C range. This sensor was chosen because does not require any external calibration or trimming to provide its typical accuracy.

Figure 1 shows the implementation of a Wireless Sensor Network (WSN) based on IEEE 802.15.4, Zigbee and communicate wirelessly with the sensor NI WSN-3202 for the 3 sensors through MDA300 data acquisition system. These wireless sensors communicate with the memsic wireless base station from Crossbow, which is programmable with the Mote-VIEW Software, can communicate with NI wireless sensor network (WSN) devices. The network is scalable up to many WSN nodes (in a mesh topology); having also the features of dual Ethernet ports to provide flexible connectivity to other devices in your measurement system, such as enterprise-level networks or wired I/O systems. With this flexibility, we can configure this network according our application to monitoring and measurement of environmental parameters which effected in the fabrication of semiconductor industries to prevent and malfunction and to increase the yield of the production.

Each node connects with a smart sensor, namely: a dust detection device or temperature or humidity, transducer interface model (TIM) and Network Capable Application processor (NCAP), as shown in Figure 1.

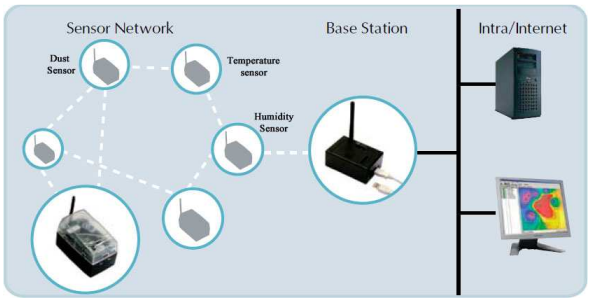


Figure 2. Environmental monitoring network.

 ISBN: 978-1-61208-203-5

The server acquires the monitoring information from the distributed network of smart sensors and processes this data via specialized software. Based on the user configured thresholds, the server will either take no action, but to record the data for statistic purposes, or send a signal to other devices for specific tasks, such as air trap shutdown or activating air recirculation systems, in case the configured thresholds have been surpassed to a critical level. This depends on the specific application for which the dust detection sensor or humidity or temperature network is used.

The network consists from 3 nodes via data acquisition MDA300 with NI WSN-3202 for dust, temperature and humidity. Figure 3, depicts the practical picture of the NI WSN 3202 and MDA 300.

Figure 3. Real-life pictures of used equipment.

The NI WSN-3202 measurement node is a wireless device that provides four ±10 V analog input channels and four bidirectional digital channels that we configured on a per-channel basis for input, sinking output, or sourcing output. The 18-position screw-terminal connector delivers direct connectivity to sensors and offers a 12 V, 20 mA sensor power output that use to drive sensors that require external power. The power for the measurement node is similar to NI WSN - 3226 (four 1.5 V, AA alkaline battery cell).

A 2.4 GHz radio wirelessly transmits data to the WSN gateway, where you can connect through Ethernet to other network devices. NI-WSN software delivers easy network configuration in NI Measurement & Automation Explorer (MAX) and data extraction with NI Mote-VIEW software. The nonprogrammable WSN-3202 does not include a license to target and program the node with the Mote-VIEW Wireless Sensor Network (WSN) Module Pioneer.

IV. MONITORING APPLICATION OF THE WIRELESS SENSOR NETWORK

The network of wireless sensors in Figure 2 is made out of wireless sensors network (WSN) based on IEEE 802.15.4. Zigbee and the acquired of data from the sensors are periodically read with a selectable multiplexing according NI Mote-VIEW software. The performance of the proposed measurement and monitoring systems depends mainly on the sensors that are used to acquire the environmental data. The resolution and conversion time of the (MDA300 data acquisition board) analog to digital converter that was used in the application is sufficient for the proposed application and can be expendable for using with many sensors. The Topology view shows a map of the network of Motes, including placement and parenting information. This allows the user to define and view a topology of their Mote deployment.

The front panel of the application used for monitoring and measurement of environmental parameters is presented in Figure 4. As it can be observed, each signal from the sensor is displayed. The application allows the user to set the variation limits for every channel.

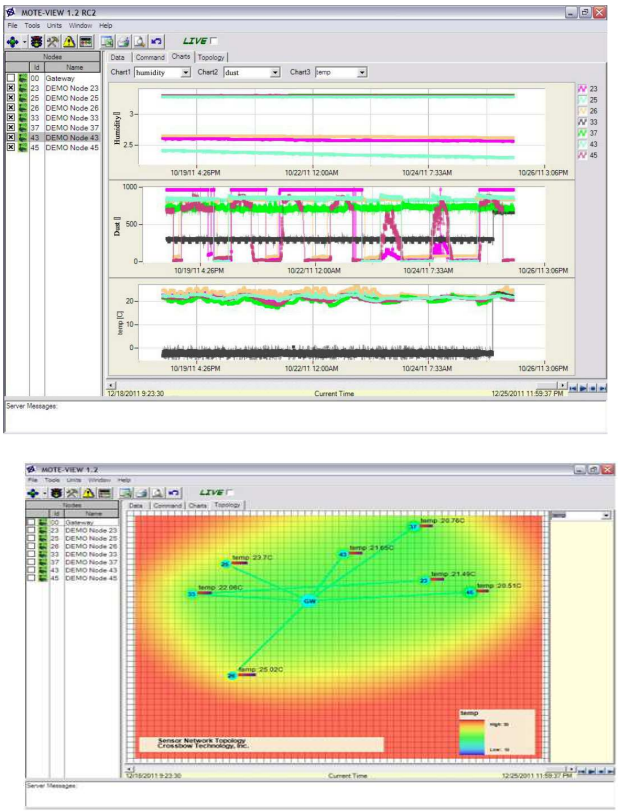


Figure. 4 The interface of the application used to display the acquired data from sensors

The experimental results were obtained by using several types of dust: sand dust with high granularity, plaster dust and smoking ash. Another dust detector has been used as a reference, based on the gravimetric principle "D-RC80 Automatic sampling device for Gravimetric Dust measurements", used as reference measuring system. The output of the sensor is sent through the MDA300 based on Mote_view software to the server.

For the smoking ash, we obtained a fluctuation in the results, as shown in Figure 5, with a solid average, which was within the values obtained by using the dust detector with the gravimetric principle.

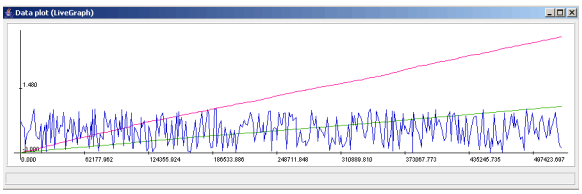


Figure.5 Graph of smoking ash measurements.

The experimental results depicted in Table I are encouraging regarding the accuracy of the optical measurement, compared to the ones made with a gravimetric device. The mean values were calculated based on 20 measurements.

TABLE I. MEAN EXPERIMENT RESULTS

Type of dust	Mean measurement with our setup	Gravimetric measurement
Sand	3.0 mg/s	3.4 mg/s
Plaster	2.9 mg/s	3.9 mg/s
Smoking Ash	1.43 mg/s	1 mg/s

We also conducted tests with other two sensors for humidity and temperature, that were sending data simultaneously to the server and we obtained satisfactory average results, having the humidity around point 27% and the temperature around 22C°.

Based on the user configured thresholds, the server will either take no action, but to record the data for statistic purposes, or send a signal to other devices for specific tasks, such as air trap shutdown or activating air recirculation systems, in case the configured thresholds have been surpassed to a critical level. This depends on the specific application for which sensor network is used.

V. CONCLUSIONS AND FUTURE WORKS

Environmental parameters (humidity, dust, temperature, etc.) have significant effects on electronic fabrication and especially nanoelectronic industries modules and circuits, those been necessary to be continuous monitored and measured, especially in industrial areas. Each parameter sensing device can focus on a specific area and by managed as a single entity or in turn it can be used as only one point of presence in an area, contributing to the overall accuracy of the measurement. The human interaction will be greatly reduced by using such a network. Also, compared to human observation, the introduction of a smart sensor network is more flexible when it comes to dangerous and hostile environments where humans can't penetrate, allowing access to information previously unavailable from such close proximity. Future work aims at improving the performance and durability of smart sensors networks and to prove them as a versatile application. We also aim to increase the ability of the dust sensors to make discrimination regarding the type of dust and based on this to configure the system's threshold value in order to make a decision according to the settings and applications.

Sensor scheduling can be obtained by enabling the sensor nodes to modify communication requirements in response to network conditions and events detected.

From the experience of already existing devices, we can expect that in the coming decade a large number of monitoring systems for all physical phenomena will emerge, with great application in the human health sector, industrial sector and the environment. The monitoring system gives excellent opportunities to design and configure many types of sensors to monitor and control all physical phenomena for many applications based on people demands.

REFERENCES

[1] S. Franssila, "Introduction to Microfabrication", 1st ed, John Wiley & Sons, 2004, ISBN 0-470-85106-6.

[2] Oliver Geschke, Klank & Telleman, eds, "Microsystem Engineering of Lab-on-a-chip Devices", 1st ed, John Wiley & Sons. ISBN 3-527-30733-8.

[3] N. P. Mahalik, "Micromanufacturing and Nanotechnology", Springer, 2006, ISBN 3-540-25377-7.

[4] IEEE Instrumentation and Measurement Society "IEEE 1451.5, Standard for a Smart Transducer Interface for Sensors and Actuators– Wireless Communication and Transducer Electronic Data Sheet (TEDS) Formats", TC-9, The Institute of Electrical and Electronics Engineers, Inc., New York, N.Y. 10016.

[5] K. Weide-Zaagea, W. Horaudb, H. Frémont, "Moisture diffusion in Printed Circuit Boards: Measurements and Finite- Element-Simulations", 16th European Symposium on Reliability of Electron Devices, Failure Physics and Analysis, Microelectronics Reliability, Vol. 45, pp. 1662-1667, 2005.

[6] Sheet No.: E4-A01501EN GP2Y1010AU0F Compact Optical Dust Sensor, SHARP Corporation, 2006.

[7] Mote-VIEW 1.2 User's Manual Revision B, January 2006, Crossbow Technology, Inc.

[8] T. Defeng, L. Shixing, X. Wujun, and Z. Yongming "A Fire Monitoring System in ZigBee Wireless Network" CYBERC 2010 - The 2nd International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery CyberC 2010. October 10-12, 2010. Huangshan/China.

[9] L.B. Ruiz , J.M. Nogueira, and A.A.F. Loureiro, "MANNA:A Management Architecture for Wireless Sensor Networks" IEEE Communications Magazine, vol. 41, no. 2, pp. 116–125, 2003.

[10] Data Sheet of Humidity Sensor HIH 3605, Honeywell Corporation.

[11] Data Sheet of Temparature Sensor LM35, National Semiconductors.

[12] Data Sheet of MDA300, Data Acquisition Board, Crossbow Technology Inc.

[13] Data Sheet of Memsic Starter Kit, WSN-START, Crossbow Technology Inc.

[14] Data Sheet of NI WSN 3202, the Programmable Measurement Node, National Instruments.

 ISBN: 978-1-61208-203-5

Semi-Blind Dual-Hop Relay Selection based on Long-term Channel Statistics on the First Relaying Hop

John F. An, Sheng Yang and Sheng Yuan Wu
Department of Communications, Navigation and Control Engineering Department
National Taiwan Ocean University, Keelung, Taiwan
Republic of China

Abstract—**In this paper, we propose a novel relay selection scheme based on the *maximum a posteriori* (MAP) decision criterion for a dual-hop semi-blind (DHSB) amplify-and-forward (AF) relaying, where only long-term channel statistics at the first relaying hops are taken into consideration. The observed channel gain vectors across the cluster of the first hops are provided to the selection algorithm, which then selects the relay node which has the "*most likely highest probability of channel power gain distribution*" using a posteriori likelihood ratio over a burst time period. The other selection scheme based on the maximum first-hop signal-to-noise power ratio (SNR) is also compared, taking into account sample mean of channel power gain. Fixed gain relaying is adopted to constrain our DHSB system for both MAP and first-hop SNR based relay selection schemes. The outage performance of max-min sense end-to-end SNR relay selection, as a conventional scheme and benchmark, is compared to our proposed schemes. Simulation results demonstrate that, in terms of outage performance, our proposed MAP-based relay selection scheme is particularly appropriate for a DHSB relaying network.**

Keywords- Relaying network; dual-hop; outage probability; maximum a posteriori probability; Rayleigh fading.

I. INTRODUCTION

Various selective cooperation schemes have been investigated recently with the objective of improving transmission throughput and reducing outage probability [1, 2, 3, 4]. There are two common selection relay schemes for which the amplify-and-forward (AF) and decode-and-forward (DF) modes are adopted. For practical relay networking, it is realized that the AF relaying protocol provides the simplest and most economic relaying approach [1, 2, 5, 6, 7], where the relay node (R) forwards the received signal from the source (S) to destination (D) after scaling it to meet its power constraint. The relay selection schemes presented in the literature [1, 5], using the so-called max-min relay selection scheme, are primarily based on the selection criteria of instantaneous SNR across both relaying hops (S-R $\cap$ R-D), with the proviso that the achieved transmission rate is satisfied. However, this approach likely introduces high computational complexity of the relay link instantaneous channel power gains (via channel estimations) and inadequate real-time capability for executing the selection algorithm [8, 9]. As a consequence, it results in performance degradation of the relay selection operationally. These drawbacks can be mitigated if relay selection solely depends on the long-term channel statistics of the source to relay link (i.e., first-hops), where the dual-hop semi-blind (DHSB) AF relay was introduced [12, 14].

In reality, it is impossible to switch (select) the relay node per symbol base, but which can be determined objectively in terms of sample mean of SNR_i with long-term channel statistics. Therefore, the maximum a posteriori probability (MAP) is based on the maximum-likelihood decision criterion which is simple and needs a minimal amount of statistical information (i.e., probabilistic channel description). Similar to the MAP approach using long-term channel statistics, a first-hop SNR-based relay selection is also considered in the DHSB relay system, in comparison with MAP-based outage performance. This is measured on the first and second statistical moments of the channel parameters (i.e., sample mean of channel power gains) without stochastic channel description. However, the conditional mean is not linear in the time-varying fading channel. Therefore, it is possible to cause a large error variance of the channel power gain that will result in the performance degradation on the relay selection. Accordingly, that gives the instantaneous end-to-end SNR based max-min relay selection a degraded performance. Hence, we turned our attention to the MAP relay selection, and the problem in determining the optimum channel range over the observed relaying links which has the minimum average risk for the relay selection. We assumed that the perfect channel state information (CSI) and statistics orders are estimated at the source node via the down-link pilot sub-channels. These parameters are then fed back to the destination for decision algorithm via dedicated uplink control channel sequentially.

Recently, a performance analysis of DHSB AF relaying over Nakagami-*m* fading channel has been investigated based on the end-to-end SNR [11, 14], however, there are no relay selection methods being discussed. In this paper, we have focused on the outage performance of a DHSB AF relaying scenario with our extended the maximum a posteriori probability (MAP) decision algorithm for the relay selection. Hence, the MAP-based relay selection scheme does not require calculation and comparison of end-to-end signal-to-noise ratio (SNR) across relay hops, since it calculates the a posteriori probability using long-term channel statistics and then selects the relay link with the highest probability over multi-relaying links. The main contributions of our work can be summarized as follows.

A) A novel relay selection scheme based on MAP decision rule is proposed for the DHSB relay system, where M-1 likelihood ratios of the first hop channel gains are exploited. Hence, our proposed selection algorithm makes $\frac{1}{2}M \times (M-1)$ comparisons, instead of

 ISBN: 978-1-61208-203-5

M^M comparisons for the general max-min end-to-end SNR relay selection scheme. This greatly reduces the complexity of implementing the selection

B) Our presented outage probabilities also include the performance constraint on the relay selection where it was not discussed previously in most of the publications [1, 5, 9, 12].

The rest of this paper is structured as follows. Section II describes the system model and AF relaying implementation. Section III discusses implementation issues of the MAP relay selection scheme. Simulation and analytical results are compared in terms of outage probabilities are provided in Section IV. Section V concludes this paper.

II. SYSTEM MODEL

Fig. 1 shows a dual-hop AF relaying network incorporating the MAP decision algorithm. For the m^{th} relay node, m = 1, 2,…, M, the channel gains, $r_{S,Rm}$ and $r_{Rm,D}$, denote the first hop from the S node to the R node and the second hop from the R node to the D node respectively. We assume that independent and identically distributed (i.i.d.) static Rayleigh fading [10] occurs across all relay hops S-R and R-D. The channel statistics (i.e., mean, channel covariance) of both relaying segments corresponding to the channel state information (CSI) are assumed to be perfectly estimated via the pilot sub-channels, and centralizes these parameters available to the MAP decision algorithm at D node. Hence, the AF gain can be formulated to an inverse function of the average channel power gain of the first hop [11, 12]. In our paper, the MAP decision algorithm (i.e., a special case of the Bayes decision rule) [13] is implemented to minimize the average cost per decision of relay selection, where the most likely highest probability of channel power gain distribution is measured over M first-hops on a per burst basis. The problem with minimizing the average cost is solved by selecting optimum channel gain regions.

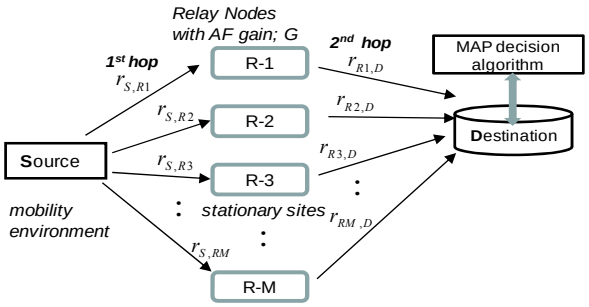


Fig.1. Dual-hop AF multi-relay networking with MAP relay selection algorithm

These optimum regions are then provided to calculate the a posteriori probabilities for each relay node. As a result, the selected relay node with the maximum a posteriori probability is determined from among the set of relay nodes.

As for the transmissions from source to relay, the instantaneous SNR on the S-R link is proportional to its corresponding channel power gain r^2_{S,R_m}, which is usually affected by the path loss (i.e. $d^{-\alpha}_{S,Rm}$), where $d_{S,Rm}$ is defined as the distance between the m^{th} relay and the source node and α represents the path loss factor. Thus, the instantaneous SNR received at relay node with transmitted source power, E_S, and white noise power, N_o, is given as

$$SNR_i = \frac{E_S}{N_o}|h|^2 d^{-\alpha}_{S,Rm} \quad (1)$$

where $r^2_{S,Rm} = |h|^2 d^{-\alpha}_{S,Rm}$ is defined as a channel power gain with respect to the channel response, h. We also assumed that all links are reciprocal.

It is assumed that instantaneous channel power gains $r^2_{S,Rm}$ and $r^2_{Rm,D}$, are independent, exponentially distributed (i.i.d.) random variables, corresponding with channel gains, $r_{S,Rm}$ and $r_{Rm,D}$ with means, $\bar{r}_{S,Rm}$ and $\bar{r}_{Rm,D}$, for m = 1, 2,…, M, respectively, according to Rayleigh fading assumptions. As a long-term post processing is applied to our relay selection scheme, the mean channel gains $\bar{r}_{S,Rm}$ and $\bar{r}_{Rm,D}$ are given by averaging over a time-slot period. Thus, the instantaneous SNRs, $\chi_1 = \sigma_1^2 |r_{S,Rm}|^2$ and $\chi_2 = \sigma_2^2 |r_{Rm,D}|^2$ are given for the first relay hop and the second relay hops, respectively, which correspond with the average SNRs, σ_1^2 and σ_2^2. It is noted that these channel gains can be obtained by estimating the CSIs via pilot sub-signals, and have Rayleigh distribution. χ_1 and χ_2 have exponential distribution (i.e., chi-squared with two degree of freedom).

Since we consider a dual-hop AF relaying system, the transmission period from the S-node to the D-node is divided into two consecutive phases. In the first phase, the S-node transmits signal s to the m^{th} R-node. Accordingly, the received signal at m^{th} R-node is given as,

$$y_{S,R_m} = r_{S,Rm} \cdot s + n_{S,Rm} \quad (2)$$

where $r_{S,Rm}$ is the channel gain between the m^{th} R-node and the S-node, and $n_{S,Rm}$ is the additive white Gaussian noise (AWGN) at the m^{th} R-node with zero mean and variance σ^2_{Rm} and the signal energy $E[ss^*] = E_S$. In the second phase, the S-node is on standby and the received signal $y_{S,Rm}$ at the R-node is amplified by a fixed gain, G, and then transmitted to the D-node. The received signal at D-node is therefore given by

$$y_{Rm,D} = G \cdot r_{Rm,D}(y_{S,Rm}) + n_{Rm,D} \quad (3)$$

where $r_{Rm,D}$ and $n_{Rm,D}$ are the channel gain between the m^{th} R-node and D-node and the AWGN at the D-node with zero

 ISBN: 978-1-61208-203-5

mean and variance σ_D^2 respectively. Since our objective for this paper is to exploit a new relay selection scheme and investigate its outage performance, the end-to-end SNR calculated for the outage evaluation will not be further discussed here. The resulting end-to-end SNR (S-R-D) via m^{th} relay node is given by [11, 12, 14]

$$\chi_3 = \frac{\left|G \cdot r_{S,Rm} r_{Rm,D}\right|^2 E_S}{\left|G \cdot r_{Rm,D}\right|^2 \sigma_1^2 + \sigma_D^2} = \frac{\chi_1 \chi_2}{\chi_2 + C} \qquad (4)$$

where C is a constant for AF fixed relay gain

$$C \cong \frac{E_S}{G^2 \cdot \sigma_R^2}.$$

The square of relay gain is given by [12 , 14]

$$G^2 = \frac{E_S}{\sigma_1^2 \left(\overline{\chi}_1 + 1\right)} \qquad (5),$$

where $\overline{\chi}_1$ is the mean power gain of χ_1.

In the following, we consider that the relay system operates in a half-duplex mode (i.e., time division duplexing system) and only one selected relay node is allowed to transmit per time slot.

III. MAP Relay Selection Scheme

A. Relay Selection Criterion

For a DHSB AF relaying system, the relay selection criterion will proceed to jointly search for the channel statistics which has the maximum a posteriori probability (MAP) over the first hop (stochastic fading channel). The selected relay has $\chi_{S,Rm \cup Rm,D} > 2^{2\nu} - 1$ after scaling it to meet its transmit power constraint for a target spectral efficiency ν (bit/sec/Hz). As such, a relay selection rule can be classified as M-1 likelihood ratios for each relay link (first hop between S-R), and this is subject to : the vector observations of the channel gains, the conditional probability density function, a prior probability and decision cost factor with respect to each radio link. Therefore, the a posteriori probabilities can be determined in terms of these given information individually using Bayes's rule [13]. The MAP selection process can thus be divided into two steps: (1) jointly identify the optimum channel gain ranges on the first hops where the average decision risk is minimized; and (2) achieve maximum a *posterior* probability by integrating its corresponding channel gain range. Let the conditional PDF, $P_{m1}\left(\overline{r}/L_{S,Rm}\right)$ of the channel gain vectors $\overline{r}$ $\sim \left[0 < r \le 3\sigma_0\right]$ be known at D node, m =1, 2, .. , M, and these channel gain vectors are random processes with Rayleigh distribution and its variance $E\left[r_{S,Rm}^2\right] = \sigma_0^2$. The channel gain vectors are composed of the channel sequences of the first relay hops. This gives our analysis a boundary over three times the standard deviation of the Rayleigh distribution, $\left[0 \le \Re_{11} \cup \Re_{21} \cup \Re_{31} \cup .. \Re_{M1} \le 3\sigma_0\right]$ for each first relay hop. It also corresponds to the probability of exceeding the Rayleigh envelope by one percent (1%) via setting the standard deviation $\sigma_0 = 3$. These Rayleigh fading channels were generated using Jakes' model [10]. For the m^{th} selected relaying link, $L_{S,Rm}$, the MAP-based relay selection criterion is given by

$$\hat{\Omega} = \arg \max_{r \in R_{m1}} \left\{ \int_{\Re_{m1}} \mathrm{P}_{r_{m1}}\left(\overline{r} / L_{S,Rm}\right) \mathrm{d}\, r \right\} \qquad (6),$$

m = 1, 2, .., M,

where $\Re_{m1}$ is the optimum region of the channel gain w.r.t. the first hop of the m^{th} relay node, and the conditional probability density function (PDF), $\mathrm{P}_{r_{m1}}\left(\overline{r} / L_{S,Rm}\right)$, has a Rayleigh distribution. The integrand term inside the bracket of (6) is described as a *posteriori* probability of the channel power gain distribution w.r.t. optimum region R_{m1} the first relay hop $r_{S,Rm}$, and is given by

$$P\left(L_{S,Rm} / R_{m1}\right) = \int_{\Re_{m1}} P_{r_{m1}}\left(\overline{r} / L_{S,Rm}\right) dr \qquad (7)$$

where m =1, 2, .. , M and $\sum_{m=1}^{M} P\left(L_{S,Rm} / R_{m1}\right) = 1$.

Those $\Re_{m1}$ are determined by the MAP decision algorithm (which will be discussed in next sub-section). Hence, the m^{th} relay node selection is determined if

$$P\left(L_{S,Rm} / R_{m1}\right) > P\left(L_{S,Rk} / R_{k1}\right) \qquad (8)$$

for all k ≠ m, and by denoting the maximum decision factor $\beta_m = P\left(L_{S,Rm} / R_{m1}\right)$ for simplicity of notation. It should be noted that our proposed MLD-based relay selection algorithm makes $\frac{1}{2} M \cdot (M-1)$ comparisons, instead of M^M comparisons for the max-min sense based relay selection schemes [7]. In fulfilling the DHSB AF relaying system design, we found that this simplified selection rule is more practical since it resulted in faster selection by eliminating the search through all possible end-to-end SNR comparisons in general relay selection schemes.

B. MAP Decision Algorithm

Consider an extended Likelihood Decision algorithm (Bayes decision rule) for 1-by-M multiple relay links (S-Rs) over the *M*-likelihood of receiving relay nodes, $L_{S,R} = \left[L_{S,R1}, L_{S,R2}, .., L_{S,Rm}, .., L_{S,RM}\right]$ represent the radio link vector corresponding to the first hops w.r.t. M relay nodes. For minimization of the average decision risk per relay selection using Bayes's rule [11]

 ISBN: 978-1-61208-203-5

$$P\left(L_{S,Rm} / R_{m1}\right) = \frac{P_{r_{m1}}\left(\bar{r} / L_{S,Rm}\right) \cdot P(L_{S,Rm})}{P(r)} \quad (9),$$

and the *average risk* for a selection decision is defined as [11]

$$\hat{C} = \sum_{k=1}^{M}\sum_{m=1}^{M} P\left(deciding\, L_{S,Rk} / L_{S,Rm}\right) \cdot P\left(L_{S,Rk}\right) \cdot C_{L_{S,Rk},L_{S,Rm}} \quad (10)$$

where the a *priori* probability of each relay link, $P\left(L_{S,R1}\right), P\left(L_{S,R2}\right), ..., P\left(L_{S,RM}\right)$, is equal (i.e. 1/M) and the cost factor, $C_{L_{S,Rk},\, L_{S,Rm}}$ is associated decision of classifying a link from $L_{S,Rk}$ given that it is from a link $L_{S,Rm}$. $P\left(deciding\, L_{S,Rk} / L_{S,Rm}\right)$ is the conditional probability of deciding radio link $L_{S,Rk}$ given at that $L_{S,Rm}$ belongs, and can be further interpreted as

$$P\left(deciding\, L_{S,Rk} / L_{S,Rm}\right) = \iint \cdot\cdot \int_{\Re_{k1}} P\left(\bar{r} / L_{S,Rm}\right) dr \quad (11),$$

where $k = 1,2,..,M$ and $k \neq m$.

Note that $\Re_{k1}$ represents the optimum channel gain region at the first hop link of the k^{th} relay node. The problem is to select the optimum channel gain regions of $\Re_{11}, \Re_{21}, ..., \Re_{M1}$ such that the average selection risk (9) is minimized. Substituting (11) into (10) and separating out the costs with identical indices, then (10) can be rewritten in terms of M integrals as

$$\begin{aligned} \hat{C} &= \sum_{k=1}^{M}\sum_{m=1}^{M} P\left(L_{S,Rm}\right) C_{L_{S,Rk},L_{S,Rm}} \iint \cdot\cdot \int_{\Re_{k1}} P\left(\bar{r} / L_{S,Rm}\right) dr \\ &= \iint \cdot\cdot \int_{\Re_{11}} \sum_{m=1}^{M} C_{L_{S,R1},L_{S,Rm}} P\left(\bar{r} / L_{S,Rm}\right) P\left(L_{S,Rm}\right) dr + \\ &\quad \iint \cdot\cdot \int_{\Re_{21}} \sum_{m=1}^{M} C_{L_{S,R2},L_{S,Rm}} P\left(\bar{r} / L_{S,Rm}\right) P\left(L_{S,Rm}\right) dr + \cdots \\ &\quad \iint \cdot\cdot \int_{\Re_{M1}} \sum_{m=1}^{M} C_{L_{S,RM},L_{S,Rm}} P\left(\bar{r} / L_{S,Rm}\right) P\left(L_{S,Rm}\right) dr \end{aligned} \quad (12)$$

Without loss of the generality, a new function, $y_k(r)$, within the integrands of (12), is given as

$$y_k(r) \cong \sum_{m=1}^{M} C_{L_{S,Rk},L_{S,Rm}} P\left(\bar{r} / L_{S,Rm}\right) P\left(L_{S,Rm}\right) \quad (13)$$

k =1, 2, .. , M.

From (13), we see that the cost function (12) will be minimized if the optimum region is determined as follows: $r \in \Re_{m1}$ if $y_m(r) < y_k(r)$ *for all k = 1, 2,..., M* and $k \neq m$ (i.e., likelihood ratio). Each of the optimum channel gain regions is found by taking M-1 comparisons of $y_m(r) < y_k(r)$ for the m^{th} first relay hop, where the intersection of M-1 channel ranges, is determined accordingly

$$\Re_{m1} = \bigcap_{\substack{k=1,2,..,M \\ m \neq k}} \left(y_m(r) < y_k(r)\right) \quad (14)$$

These mutually exhaustive and exclusive decision regions, $\Re_{11}, \Re_{21}, .., \Re_{M1}$, make the average cost, $\hat{C}$, that is minimized. From (5), it is clear that the MAP decision rule for the relay selection is obtained using a zero-one cost assignment [i.e., cost factor $C_{L_{S,Rk},L_{S,Rm}} = 0$ for k = m, and $C_{L_{S,Rk},L_{S,Rm}} = 1$ for $k \neq m$] and an equal a priori probability of each relay link.

Fig. 2 shows one of the simulation results of optimum regions generated from (14), where the optimum regions are R_{11} = [0.01~0.95], R_{21} = [1.65~3.0], R_{31} = [1.29~1.64] and R_{41} = [0.96~1.28]. During simulations, maximum Doppler frequency, f_d, exists at all radio links of the first hop with normalized $f_d T = 0.05$ and data duration T. This simulation is performed using four i.i.d. Rayleigh fading channels [10]. By using these optimum regions, the a posteriori probabilities are obtained as $P\left(L_{S,R1} / R_{11}\right) = 0.403$ $P\left(L_{S,R2} / R_{21}\right) = 0.3016$, $P\left(L_{S,R3} / R_{31}\right) = 0.1681$ and $P\left(L_{S,R4} / R_{41}\right) = 0.1141$ respectively. Note that the channel covariance of radio link $L_{S,Rm}$ for each conditional probability is obtained by averaging 100 channel observations, respectively. Accordingly, the 1st relay node has the MAP = 0.403, and is selected as the "best" relay in accordance with our MAP relay selection criterion (6).

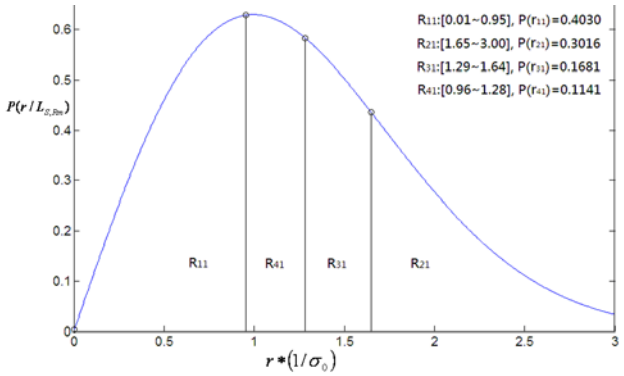


Fig. 2 Channel power gain distribution with the first hop channels are characterized by normalized Doppler frequencies $f_D T_S$ =0.05 and standard deviation $\sigma_0 = 3$ for M = 4.

IV. NUMERICAL ANALYSIS

In this section, we illustrate by simulation the outage performance of our proposed MAP-based relay selection scheme on behalf of a DHSB AF relaying system. Our analyses started with relay selection using the MAP decision algorithm, then the outage performances were analyzed via Monte-Carlo simulations. For a single relay link, the outage is generally given by $P_{out} = \mathrm{P_r}\left(\chi_3 < 2^{2v} - 1\right)$, where χ_3 is defined in (4).

 ISBN: 978-1-61208-203-5

The target spectral efficiency is υ = 1 bps/Hz and M = 4. To fulfill our optimum decision input parameters and numerical analyses, the variances of the instantaneous channel gain of the first hops were calculated in the basis of block lengths of 100 channel samples at each relay link and the outage probabilities were averaged over a collection of 2000 channel segments for each SNR value. Note that the relay selection is performed form block to block and the average SNR is considered in symmetric relay hops (i.e., $\sigma_1^2 = \sigma_2^2$ single hop SNR). The relay links are generated using Rayleigh fading channel model [10] with normalized Doppler frequency $f_d T = 0.05$, and have equal average end-to-end channel power gain across all relay links.

In Fig. 3, the outage performance of MAP-based relay selection (DHSB AF), is compared with the analytical max-min based selection [Equ (5), 5] as a benchmark, and the first-hop SNR based selection (semi-blind AF) are also presented. Both MAP and first SNR based relay selection schemes are provided for the DHSB AF relay networking. The simulations were conducted for multi-relay M = 2, 3, 4, respectively. We found that with an increase in the number of relay nodes, the outage probabilities decrease as selection diversity gain is available correspondingly in all cases. Comparing with the max-min sense analytical results, our proposed MAP scheme introduces about 3 dB degradation at a $10^{-3.}$ outage rate, whilst there is a 4 dB degradation with first hop SNR-based selection. It is generally accepted that end-to-end SNR selection gives higher expected reward, but will consume more computational load.

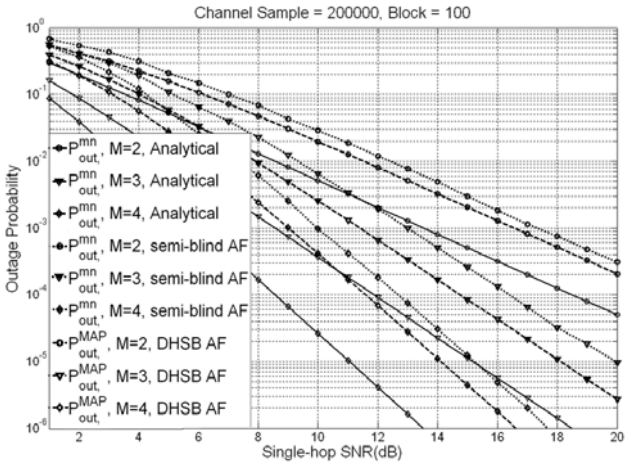


Fig.3. Outage performance comparison with max-min end-to-end SNR based selection (Analytical) and proposed MAP-based selection scheme (DHSB AF) and first-hop SNR based (semi-blind AF) with multi-relay M = 2, 3, and 4 respectively.

Through our simulation results, we confirmed that our proposed MAP-based relay selection scheme consumes approximately 1 dB less energy than the first hop SNR-based selection when considered with a DHSB AF relaying system. Although a full-CSI assisted end-to-end (max-min) selection scheme has better performance than DHSB AF relaying, this incurs a greater computational load in selection algorithm implementation. As a result, our proposed scheme gives a simple and effective approach to practical DHSB AF relay networking design.

V. CONCLUSION AND FUTURE WORK

Through our simulations, we confirmed that the MAP-based probabilistic channel description approach relay selection outperforms the first-hop sample mean SNR based selection for a DHSB AF relaying system using a long-term channel statistics scenario. We also found that it introduces a 3 dB performance degradation at outage level 10^{-3} against max-min sense relay selection scheme, whilst substantially simplifying and reducing the relay selection process in terms of computational load. The numerical and simulation results demonstrate that our proposed MAP-based relay selection is appropriate to DHSB relay network design in terms of the implementation simplicity and outage performance. It is also interesting to derive an analytical expression of the outage probability jointly considering the relay selection performance (i.e., MAP-based selection algorithm), not merely adopting a generic outage definition. This will be helpful to evaluate the overall system performance in practical relaying network design.

REFERENCES

[1] A. Bletsas, A. Khitsi, D. P. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," IEEE Journal of Selected Areas Communications, Vol.24, No.3, March, 2006, pp. 659-672.

[2] Cheng-Xiang Wang, Xuemin Hong, et al., "Cooperative MIMO Channel Models: A Survey," IEEE Communications Magazine, February 2010, pp. 80-87.

[3] Yang Yang, Honglin Hu, Jing Xu and Guoqiang Mao, "Relay Technologies for WiMAX and LTE-Advanced Mobile Systems," IEEE Communications Magazine, Vol. 47, No.10, October 2009, pp. 100-105.

[4] Ritesh Madan, Neelesh B. Mehta, Andreas F. Molisch, ans Jin Zhang, "Energy-Efficient Cooperative Relaying over Fading Channels with Simple Relay Selection," IEEE Transaction on Wireless Communications, Vol. 7, No. 8, August 2008, pp. 3013-3025.

[5] Abdulkareem Adinoyi, Yijia Fan, Halim Yanikomeroglu, H. Vincent Poor, and Furaih Al-Shaalan., "Performance of Selection Relaying and Cooperative Diversity," IEEE Trans. on Wireless Communications, Vol. 8, No. 12, December 2009, pp. 5790 -5795.

[6] Chris Conne and Il-Min Kim, "Outage Probability of Multi-hop Amplify-and-Forward Relay Systems," IEEE Transactions on Wireless Communications, Vol. 9, No.3, March 2010, pp. 1139-1149.

[7] Elzbieta Beres and Raviraj Adve, "Selection Cooperation in Multi-Source Cooperative Networks," IEEE Transactions on Wireless Communications, Vol. 7, No. 1, January, 2008, pp. 118-127.

[8] Yubo Li, Qinye Tin, wei Xu, and Hui-Ming Wang, "On the design of relay selection strategies in regenerative cooperative networks with outdated CSI,"IEEE transactions on Wireless Communications, Vol. 10, No. 9, September 2011, pp. 3086-3097.

 ISBN: 978-1-61208-203-5

[9] M.O. Hasna and M. S. Alouini," A performance study of dual-hop transmissions with fixed gain relays," IEEE-ICASSP, Vol. 4, April, 6-10, 2003, pp. 189-192.

[10] Y.R. Zheng and C. Xiao, "Simulation models with correct statistical properties for Rayleigh fading channels," IEEE Transaction on Communications, Vol. 51, No. 6, June 2003, pp. 920-928.

[11] Nuwan S. Ferdinand and Nandana Rajatheva, "Unified Performance Analysis of Two-Hop Amplify-and-Forward Relaying Systems with Antenna Correlation," IEEE Transactions on Wireless Communications, Vol. 10, No. 9, September 2011, pp. 3002-3011.

[12] Mazen O. Hasna and Mohamed-Slim Alouini, "A Performance Study of Dual-Hop Transmission with Fixed Gain Relays," IEEE Transactions on Wireless Communications, Vol. 3, No. 6, November 2004, pp. 1963 -1968.

[13] Lonnie C. Ludeman, "Random Processes - Filtering, Estimation, and Detection," Wiley, 2003, pp. 424-430.

[14] Minghua Xia, Chengwen Xing, Yik-Chung Wu, and Sonia Aïssa, "Exact Performance Analysis of Dual-Hop Semi-Blind AF Relaying over Arbitrary Nakagami-m Fading Channels," IEEE Transactions on Wireless Communications, Vol. 10, No. 10, October 2011, pp. 3449-3459.

 ISBN: 978-1-61208-203-5

Performance Analysis of MIMO STBC in A High Altitude Platforms Communications Channel

Iskandar Iskandar and Albaz Rosada
School of Electrical Engineering and Informatics, Bandung Institute of Technology
Jalan Ganesha no.10 Bandung 40132 Indonesia
E-mail : iskandar@ltrgm.ee.itb.ac.id, albaz@students.itb.ac.id

***Abstract*–MIMO system recently emerged as a solution for the provision of wireless communications to improve capacity and decrease bit error rate. One of MIMO variant, which is used in this paper, is Space Time Block Code (STBC). STBC allows diversity gain using combination of spatial and time dimension without changing its bandwidth requirements. This paper presents an implementation of MIMO STBC 2x1 and 2x2 on HAPS channel with the assumption that the channel state condition is known at the receiver (perfect CSIR). HAPS channel characteristic is known to follow Ricean distribution in which it depends on its K factor. In case of HAPs, K factor also depends on its elevation angle. Using computer simulation, this paper analyzes HAPs channel performance using MIMO STBC 2x1 and 2x2 on the various of elevation angles. It is shown that MIMO STBC 2x1 and 2x2 are able to increase performance of HAPs channel including HAPs channel at low elevation angle. However from our simulation capacity improvement of MIMO STBC is obtained insignificant, therefore we propose MIMO spatial multiplexing, which is another variant of MIMO to obtain more capacity.**

Keywords– HAPs; MIMO; STBC; Ricean channel; K factor; Spatial Multiplexing.

I. INTRODUCTION

High Altitude Platforms (HAPs) is an object floating on a stratospheric layer bringing wireless communication equipment at approximately 17-22 km above the ground. HAPs is able to exploit much the advantages and at the same time overcome the drawback of the traditional systems in terms of propagation delay and path loss suffered by satellite system or a huge number of base station required by the terrestrial system.

In our previous research [1], HAPs channel characteristic which is experimentally measured in semi-urban environment, deteriorates at low elevation angle. In other word the performance of HAPs communication needs improvement for the users who are located at the edge coverage. Measurement result shows that for low elevation angle, i.e. lower than 40^0, fading depth is observed to be approximately 25 dB or more. Such huge fading depth, of course, will limit HAPs service coverage to elevation angle only higher than 40^0 or about 50 km in diameter of service coverage. To overcome such problems MIMO STBC is proposed in this work. MIMO STBC which allow diversity gain is expected to improve bad channel condition especially for low elevation angles.

MIMO is simply defined as an use of more than one antenna at transmitter and/or receiver. There are two kinds of MIMO called Spatial Multiplexing (SM) and Space Time Block Code (STBC). On this paper, we use MIMO STBC with 2 antennas transmitter with combination of 1 and 2 antennas at receiver. HAPS is then used as transmitter and both antennas is placed onboard the platform as depicted in Fig 1. The previous research [3] have shown that MIMO can be implemented on single HAPs with specific spacing between them depend on its frequency. For 2.4 GHz, both antennas must be separated about 12 meters. Simulation is then runned by MATLAB R2008a. The variables that are used in the simulation is elevation angle from 10 – 90 degree which represent the K factor and operating frequency at 1.2 and 2.4 GHz.

The remaining part of this paper is outlined as follows. Section 2 presents channel model and propagation characteristic in a HAPs system. Section 3 reviews in detail a concept of MIMO STBC. Simulation model is explained in Section 4. Section 5 shows simulation result, and finally, concluding remark is drawn in Section 6.

II. HAPS CHANNEL MODEL AND PROPAGATION CHARACTERISTIC

Generally, there are some propagation phenomena that can happen on HAPS channels as follow: Free space path loss, multipath fading, rain attenuation, gas absorption, and scintillation [2]. Most of them are frequency dependence. Rain attenuation, gas absorption, and scintillation are significant only on a high operating frequency, i.e. above 10 GHz. While this paper used freq 1.2 and 2.4 GHz, all of them will be ignored on the formulation and simulation.

In case of HAPs channel, Ricean fading is a general case of fading channel model that there are two components of signal arrive at the receiver. First component arrive at receiver

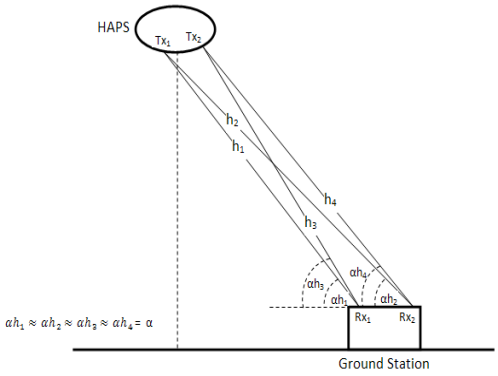


Fig. 1 MIMO STBC model in HAPs channel.

 ISBN: 978-1-61208-203-5

through line of sight (LOS) path and the second component come from multipath scattered signal. In HAPs communication channel, it is possible to have both components because HAPs is highly positioned above the ground. Therefore, the channel characteristic in HAPs system can be represented by Ricean distribution which is expressed as follow,

$$x(t) = \sqrt{\frac{K\Omega}{K+1}}\, e^{j(2\pi f_D \cos(\theta)t + \Phi} + \sqrt{\frac{K}{K+1}} h(t) \qquad (1)$$

where K is a Ricean factor, θ and Φ are elevation angle, f_D is Doppler frequency from receiver movement with velocity (v), and $h(t)$ is a scattered component that can be expressed as,

$$r_h(\tau) := E[h(t)h^*(t+\tau)] = \int_{-\pi}^{\pi} p_h(\theta)\, e^{j2\pi f_D \cos(\theta)\tau}\, d\theta \qquad (2)$$

If $E[h^2(t)]$ is estimated to be one, then scattered signal power on the formula above become σ^2. On the other hand, LOS signal power which is significant in HAPs channel, is denoted by A^2, and K is defined as $A^2/2\sigma^2$. So, the total received power is represented by

$$E[x^2(t)] = A^2 + 2\sigma^2 \qquad (3)$$

$E|x^2(t)|$ is local mean received power. Therefore it is to be said that Ricean signal is an addition of LOS and NLOS component with a weighting factor of K. Then, the formula above can be written as follow

$$H = \sqrt{\frac{K}{K+1}}.H_d + \sqrt{\frac{1}{K+1}}.H_s \qquad (4)$$

where H_d is LOS component, and H_s is NLOS component.

III. MIMO SPACE TIME BLOCK CODE

Now we simple analyze a multiple input multiple output (MIMO) technique which is defined as the use of more than one antenna at transmitter and/or receiver as shown in Fig. 2. As already mentioned, there are two kinds of MIMO called Spatial Multiplexing (SM) and Space Time Block Code (STBC). In this paper, we evaluate the use of MIMO STBC 2x2 which in HAPs channel it can be proposed to improve the user performance located at the edge of coverage. We use the transmission scheme of orthogonal STBC which is firstly introduced by Alamouti [4]. At time t, antenna T_{x0} sends signal s_0 and T_{x1} sends signal s_1, then at time t+T, T_{x0} sends signal $-s_1^*$ and T_{x1} sends signal signal s_0^*. Fig 3 shows a system configuration using MIMO STBC 2x2. Both of the signal then transmitted by two independent Rician channel h_0 and h_1. Channel is assumed same at time t+T and time t.

$$h_o(t) = h_o(t+T) = h_o = \alpha_o\, e^{j\theta_o} \qquad (5)$$

$$h_1(t) = h_1(t+T) = h_1 = \alpha_1\, e^{j\theta_1} \qquad (6)$$

Received signal is a multiplication of transmitted signal with channel and addition with AWGN noise.

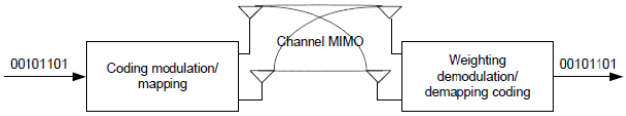


Fig. 2 Basic concept of MIMO.

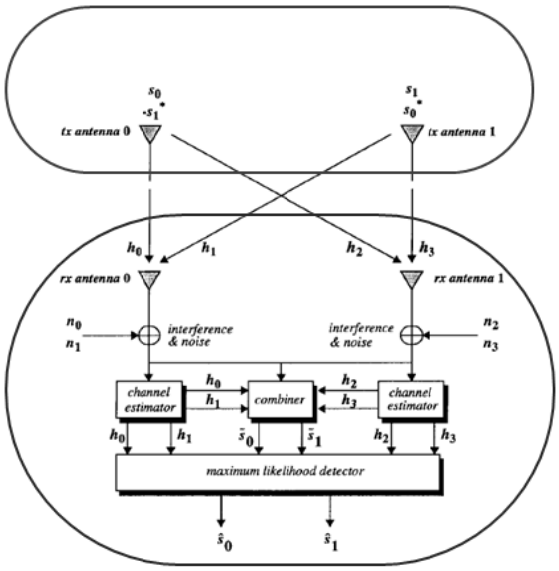


Fig. 3 MISO STBC 2x1 [4].

$$r_o = h_o s_o + h_1 s_1 + n_0 \qquad (7)$$

$$r_1 = -h_o s_o^* + h_1 s_o^* + n_1 \qquad (8)$$

$$r_{21} = h_2 s_o + h_3 s_1 + n_2 \qquad (9)$$

$$r_3 = h_2 s_1^* + h_3 s_o^* + n_3 \qquad (10)$$

Finally, to get signal s_0 and s_1, front end combiner uses channel information from channel estimator.

$$s_o = h_o^* r_o + h_1 r_1^* \qquad (11)$$

$$\tilde{s}_o = h_o^* r_o - h_1 r_1^* \qquad (12)$$

Block combiner then makes a new signal from combination of these 4 channel and 4 received signal as follows

$$\tilde{s}_0 = h_0^* r_0 + h_1 r_1^* + h_2^* r_2 + h_3 r_3^* \qquad (13)$$

$$\tilde{s}_1 = h_1^* r_0 - h_0 r_1^* + h_3^* r_2 - h_2 r_3^* \qquad (14)$$

IV. SIMULATION MODEL

The simulation model is presented in Fig 4. First, we generate random data to make a symbol stream input that consist of approximately 1000 bit. Then the data is BPSK modulated and then its output is inserted to STBC encoding block. The process that happen in this block is almost same as explained before. Bit stream is separated into two parts and

 ISBN: 978-1-61208-203-5

for the next time slot, Alamouti [4] conjugate data is sent on each antenna.

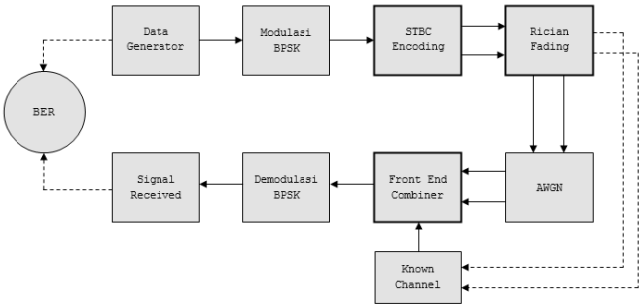


Fig. 4 MIMO STBC on HAPS channel simulation model.

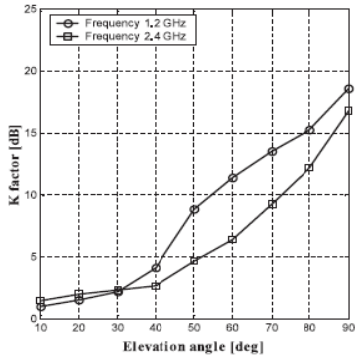


Fig. 5 K Factor and HAPS elevation angles [1].

TABLE I. SIMULATION PARAMETERS.

Frequency	1.2 GHz dan 2.4 GHz
Amount of bits	100000
Eb/No	STBC 2x1 = 0; 2; 4; 6; 8; 10; 12; 14; 16; 18; 20 STBC 2x2 = 0; 2; 4; 6; 8; 10; 12
Modulation	BPSK
Frame length	100
Number of packet	1000
Antenna Tx	2
Antenna Rx	STBC 2x1 = 1 SBTC 2x2 = 2
K Factor	Freq1.2 GHz = 0.9; 1.5; 2.2; 4.1; 8.9; 11.4; 13.5; 15.2; 18.6. Freq 2.4 GHz = 0.9; 1.5; 2.2; 4.1; 8.9; 11.4; 13.5; 15.2; 18.6.

The next process is to send the data via MIMO antenna through Ricean HAPs channel with its characteristic has been experimentally investigated in our previous work as depicted in Fig. 5. K factor as a Ricean parameter for HAPs channel has been measured and we found that its value directly governed by an elevation angle of the user that look to the HAPs [1]. Additionally, *K* factor has frequency dependency in which the higher the frequency the smaller the value of K factor. Output data from STBC encoder block in frequency domain are then multiply by this Ricean fading parameter and also added by Additive White Gaussian Noise at receiver. After that, front end combiner block processes the received data stream using channel information that in this work we assume perfect channel estimation. The extracted data from this block is then demodulated into received bit stream. This received bit stream finally compared by first bit stream sent before to get the Bit Error Rate (BER) at specific signal to

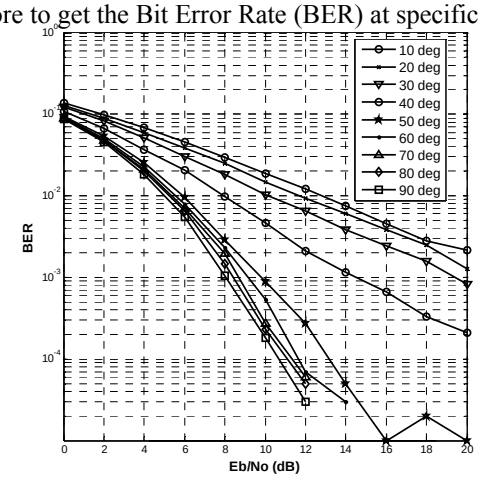


Fig. 6 Performance of SISO 1x1 (freq 1.2 GHz; elevation 0-90^0).

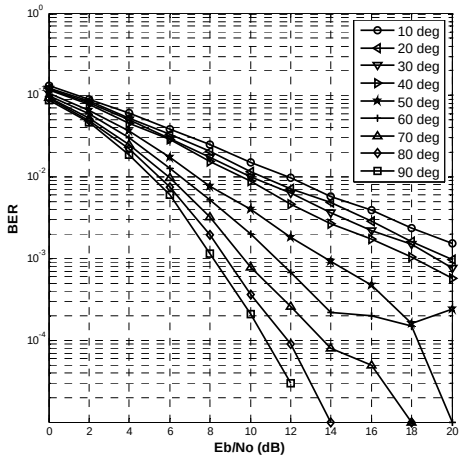


Fig. 7 Performance of SISO 1x1 (freq 2.4 GHz; elevation 0-90^0).

noise ratio (SNR) value. Table I shows parameters that is used in the simulation. We use total 100,000 bits which is separate into 1000 packet data, each of them consist of 100 bit (frame length).

V. SIMULATION RESULTS

For the comparison of all methods, SISO 1x1 HAPS system is simulated first. The result is shown in Figs. 6 and 7 for each operating frequency, 1.2 and 2.4 GHz respectively. As mentioned before, in the model, simulation is run on various elevation angles which represent K factor of Ricean channel. K Factor value is taken from previous experiment and measurement in Hokaido, Japan [1]. Based on the result, there is a gap between elevation angle 40 – 50 deg shows the profile area of K factor measurement which has a quite strong fading at low elevation angle (0 – 40 deg).

Then, the configuration is changed by adding one more antenna at transmitter creating MISO 2x1 HAPS system. Using STBC Alamouti encoding-decoding technique proposed in [3], the simulation result is shown in Fig. 8. Simulation shows that MISO can increase HAPS channel performance not only for low elevation but also for all of ground station positions. MISO is able to achieve better

 ISBN: 978-1-61208-203-5

performance of about 1-11 dB for required BER 10^{-3} variously at 10-90 deg.

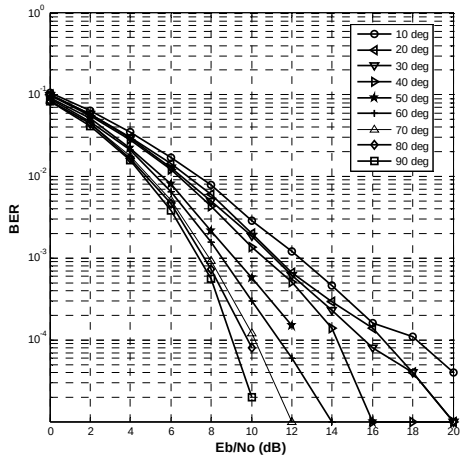


Fig. 8 Performance of MISO STBC 2x1 (freq 2.4 GHz; elevation 0-90^0).

By adding one more antenna in receiver, the configuration is now change to MIMO STBC 2x2. In Fig. 9, simulation shows that this model can make significant improvement better than the previous SISO or MISO 2x1. It has 4-17 dB improvement compared to SISO and 3-6 dB compared to MISO 2x1 at BER value 10^{-3} for all elevation angles.

By comparing all simulation results we have, it also shown that when the elevation angle is decreased (ground station is getting farther from HAPS), the impact of implementation MISO and MIMO STBC is bigger than in a high elevation angle. It can be understand because of MIMO STBC configuration is generally good for fading handling, while fading is depending on transmitter-receiver position. In a low elevation angle, fading is more severe because the path from transmitter to receiver is covered by building, trees, or another obstacle makes a NLOS signal.

While comparing MISO STBC 2x1 and MIMO STBC 2x2 on HAPS channel based on its elevation angle. At high elevation angle values, the performance of MISO STBC 2x1 is much worse than MIMO STBC 2x2, although it always still better than SISO 1x1. But at low elevation angles, the improvement is as significant as MIMO STBC 2x2.

In general, we found performances of MISO 2x1 and MIMO STBC 2x2 on HAPs channel are superior against SISO 1x1. From this results, we can also analyze the performance improvement from coverage area point of view. As mentioned before, elevation angles are so important variable in HAPS system. It can, not only represent the condition of Ricean channel, but also the coverage area itself. Fig. 10 shows the radius of coverage area of HAPs can be increased significantly from 7.28 km (SISO) to 16.78 km (MISO STBC 2x1) and 113.4 km (MIMO STBC 2x2).

Now, for the capacity analysis, we only use 3 elevation angles to make a simply understanding result (10, 40, and 90 deg). Simulation shows that MIMO STBC and MIMO Spatial Multiplexing (SM) have quite similar curve for frequency 2.4 GHz and 1.2 GHz, so that we show only result for 2.4 GHz as in Fig. 11 for MIMO STBC and Fig. 12 for MIMO SM. We can see that elevation angles have significant impact on the curve on outage capacity, but for the ergodic they have no big influences. It means that for MIMO STBC and MIMO SM, when we analyze the average HAPs channel capacity in a

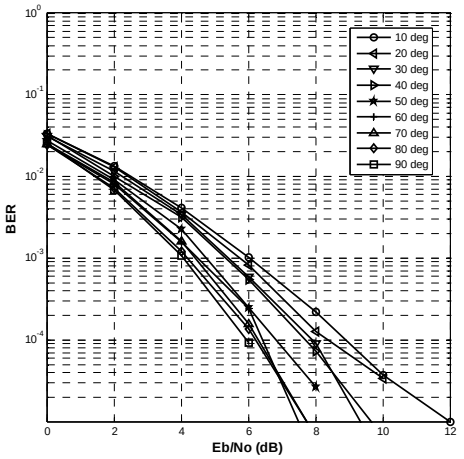


Fig. 9 Performance of MIMO STBC 2x2 (freq 2.4 GHz; elevation 0-90^0).

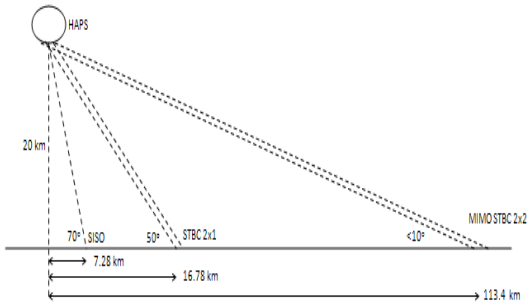


Fig. 10 HAPS radius coverage improvement by implementing MIMO STBC 2x2 and STBC 2x1.

quite long measurement time (ergodic), the elevation angles variables can be ignored, but not for outage capacity analysis.

When both curves of MIMO STBC and MIMO SM are combined together with the HAPs SISO at same elevation angle, the result is shown in Fig. 13. As we mentioned before, MIMO STBC and MIMO SM have a quite similar shape of curve, but when they are put together on one graph, its clearly shown that MIMO SM have a slope sharper than STBC. It means that MIMO SM is much more better in increasing capacity than STBC. Then SISO curve is used as a comparison (red line). We can see the line of SISO is very close together with the STBC. It means STBC is not a good method for improving HAPS channel capacity. Note that to comparing between its 3 configurations (SISO, MIMO STBC, and MIMO SM) we only use the ergodic capacity which is equivalent with Shannon capacity in SISO model.

MIMO STBC is different with MIMO SM. When MIMO SM separated two symbol stream to each antenna, STBC doesn't. As mentioned before, STBC uses Alamouti encoding which separate data in two part for each antenna Tx, then makes a "duplicate" behind them that will be sent on next time slot (t + T). For this we can understand why MIMO SM

 ISBN: 978-1-61208-203-5

can improve HAPS channel capacity extremely better than MIMO STBC.

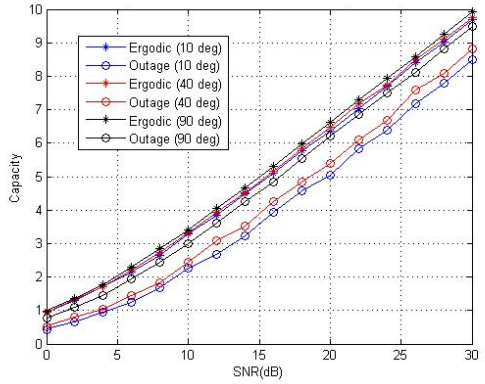


Fig. 11 Capacity of MIMO STBC 2x2 on HAPS channel (2.4 GHz).

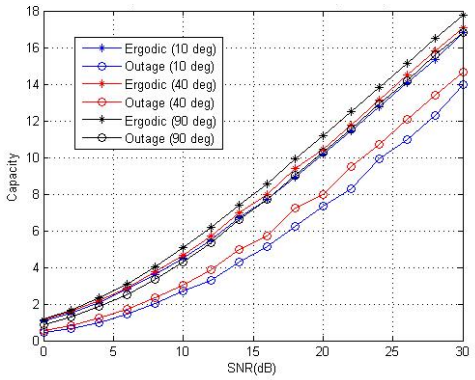


Fig. 12 Capacity of MIMO SM 2x2 on HAPS channel (2.4 GHz).

VI. CONCLUSIONS

Performance analysis of MIMO STBC on HAPS Channel has been proposed in this paper. Simulation result shows that STBC 2x1 can improve BER performance from HAPS SISO 1x1 configuration (1-11 dB on BER 10^{-3}), while MIMO STBC 2x2 even can improve more significant from SISO 4-17 dB on the same BER value for all elevation angles. By adding more antennas in receiver or transmitter may can improve a better performances. For high elevation angles (50-90 deg), the use of MIMO STBC 2x2 is much more significant than STBC 2x1. All of this improvement can also increase the radius of HAPS coverage. From 7.28 km (SISO) to 16.78 km (MISO STBC 2x1) and 113.4 km (MIMO STBC 2x2).

MIMO STBC can really improve the BER performance of HAPS system, but not the capacity. Simulation results shows that MIMO STBC have no improvement comparing to HAPS SISO configuration. To increase it, we can use another MIMO variant called Spatial Multiplexing.

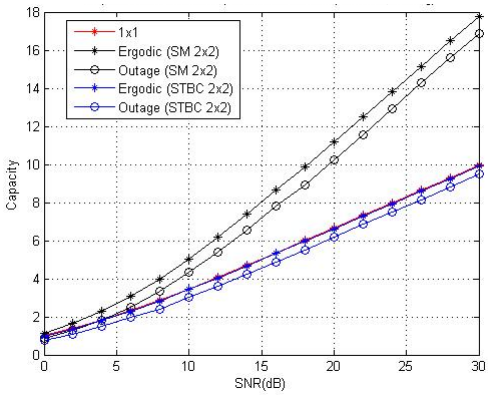


Fig 13. Capacity of SISO, MIMO SM, and MIMO STBC 2x2 on HAPS channel (freq. 2.4 GHz; elevation 90^0).

REFERENCES

[1] Iskandar, Shigeru Shimamoto, "Channel Characterization and Performance Evaluation of Mobile Communication Employing Stratospheric Paltforms", *IEICE Trans. Communication*, vol.E89-B, No.3 March ,2006.

[2] Alejandro Aragón-Zavala, José Luis Cuevas-Ruíz, and José Antonio Delgado-Penín, *High-Altitude Platforms for Wireless Communications*, 1st edition, Wiley, December 2008.

[3] E. T. Michailidis, G. Efthymoglou, and A. G. Kanatas, "Spatially Correlated 3-D HAP-MIMO Fading Channels," *International Workshop on Aerial & Space Platforms: Research, Applications, Vision, in conjunction with IEEE Globecom 2008*, New Orleans, LA, USA, December 4, 2008.

[4] S. M. Alamouti, "A simple transmit diversity technique for wireless communications", IEEE(R) Journal on Selected Areas in Communications, Vol. 16, No. 8, Oct. 1998, pp. 1451-1458.

[5] V. Tarokh, H. Jafarkhami, and A.R. Calderbank, "Space-time block codes from orthogonal designs", IEEE Transactions on Information Theory, Vol. 45, No. 5, Jul. 1999, pp. 1456-1467.

[6] A.F. Naguib, V. Tarokh, N. Seshadri, and A.R. Calderbank, "Space-time codes for high data rate wireless communication: Mismatch analysis", Proceedings of IEEE International Conf. on Communications, pp. 309-313, June 1997.

[7] V. Tarokh, H. Jafarkhami, and A.R. Calderbank, "Space-time block codes for wireless communications: Performance results", IEEE Journal on Selected Areas in Communications, Vol. 17, No. 3, Mar. 1999, pp. 451-460.

[8] M.A. Khalighi, J.-M. Brossier, and G. Jourdain, K. Raoof, "On capacity of RiceanMIMO channels," *12th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, vol. 1, pp. A150-A154, San Diego, USA, September 30-October 3, 2001.

[9] P. R. King, B. G. Evans, and S. Stavrou, "Physical-statistical model for the land mobile-satellite channel applied to satellite/HAP MIMO," in *Proceedings of the 11th European Wireless Conference*, vol. 1, pp. 198-204, Nicosia, Cyprus, April 2005.

[10] Tommy Hult, and Abbas Mohammed, "Compact MIMO Antennas and HAP Diversity for Enhanced Data Rate Communications," *VTC Spring 2007*, pp. 1385-1389, Dublin, April 2007.

[11] S. Sandu, A. Paulraj,"Space-time block codes: A capacity perspective," IEEE Communications Letters, Vol.4, No.12, December 2000.

 ISBN: 978-1-61208-203-5

Coordinated Multi-point Multistream Scheme for Disaster Recovery in MIMO Multi-Cellular Systems

Tetsuki Taniguchi Yoshio Karasawa Nobuo Nakajima
Department of Communication Engineering and Informatics
Department of Informatics
Advanced Wireless Communication research Center (AWCC)
The University of Electro-Communications (UEC), Chofu, Tokyo 182-8585, Japan
E-mail: {taniguch,karasawa}@ee.uec.ac.jp, n.nakajima@hc.uec.ac.jp

***Abstract*—Conventionally, CoMP (coordinated multi-point) transmission has been utilized for the performance improvement in the cell edge of multi-cellular systems. This paper describes CoMP scheme using multistream transmission for the disaster recovery: if base stations (BSs) lose their function by disaster attacks in some cells, user terminals (UTs) in those cells should connect to BSs in a long distance located outside their own cells under the situation where demands for the communications increase. In this case, cooperative transmission/reception of signals is considered to be helpful to keep the quality of communications. Here, our investigation is on two typical patters of allocation of cells with BS destruction, and the effectiveness of the cooperation in downlink scenario is shown through computer experiments using cooperative and noncooperative methods in the entire cell.**

***Keywords*-CoMP (coordinated multi-point); MIMO (multiple input multiple output); cooperative communication; cellular system; disaster recovery.**

I. INTRODUCTION

It is well known that CoMP (coordinated multi-point) scheme is effective for the performance improvement in the cell edge of multi-cellular systems where the signal from the target base station (BS) to user terminals (UTs) severely attenuates because of the path loss, and interferences from adjacent cells are relatively strong [1] [2]. There are many reports concerning CoMP from the viewpoint of information theoretic aspect, evaluation by computer simulation [3], measurement campaign [4], and analysis based on wave propagation [5]. All of those papers refer to the significant advantage of introducing cooperative scheme into cellular systems. Also, in our laboratory, on-computer investigation of performance is carried out in case where all of BSs, relay stations (RSs), and UTs are equipped with multiantennas, namely, multiple input multiple output (MIMO) structure.

But, when we think about the principle of CoMP, its usefulness is not restricted to the communication in peacetime – One potent candidate of application is the cooperative communication for the disaster recovery. If BSs in some cells are destroyed by an accident like landslide, UTs in those cells should connect to a BS in the next adjacent active cell, and it is normally located in a far way position. On the other hand, the traffic of disaster cell can be increase to provide their safety confirmation or to receive the further disaster and rescue information. Cooperative nature of CoMP is considered to be suitable for the improvement of this situation. From this standpoint, in this paper, a quantative evaluation of the effect of cooperation for the disaster recovery is considered based on MIMO cellular system.

The rest of this paper is organized as follows: first, Section II describes past works and what is novel in this study, and then Section III provides the system model used throughout this study and explains the simulation method. After computer simulations are carried out in Section I, Conclusions and future works are given in Section V.

II. PAST WORKS AND NOVELTY OF THIS STUDY

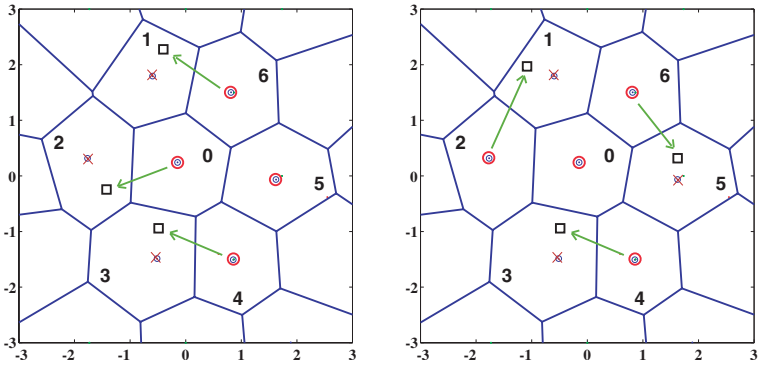
Some works have been presented as communication methods for the disaster recovery: one example is the wireless network based on virtual access point by mobile nodes in [7], and heterogeneous networks are also utilized [8]. In addition, we can easily find lots of papers concerning CoMP for the conventional use, i.e., the improvement of the cell edge performance. But, application of CoMP to disaster recovery has not been considered. We have described the primitive idea in [6] for single stream case in cell edge; it is useful as an initial investigation, but it cannot demonstrate the total ability of the system utilizing multiantenna feature. This paper extends the analysis to the multistream scenario in which UTs are located in a random position of the entire cell. In this sense, the results shown in this paper provide the tighter limit of the CoMP effect which is useful for the design of disaster-robust infrastructure.

III. SYSTEM MODEL AND DESIGN

A. Example of system model

The system is given in Fig. 1. Reflecting the fact that the uniform BS allocation is generally not possible in actual environment, the cell geometry is based on Voronoi diagram which shows the border inside which the maximum average power is derived from the BS in the same cell. As the situation before the disaster, we consider that there's one

 ISBN: 978-1-61208-203-5



(a) Pattern A (b) Pattern B

	Connection
$M_d = 1$	$UT_1 \leftarrow BS_0$
$M_d = 2$	$UT_1 \leftarrow BS_6$ $UT_2 \leftarrow BS_0$
$M_d = 3$	$UT_1 \leftarrow BS_6$ $UT_2 \leftarrow BS_0$ $UT_3 \leftarrow BS_4$
$M_d = 4$	$UT_1 \leftarrow BS_6$ $UT_2 \leftarrow BS_0$ $UT_3 \leftarrow BS_0$ $UT_4 \leftarrow BS_5$

(c) Defect cells and new connection.

Figure 1. System model. Circles, cross marks, and box marks in (a) and (b) respectively denote working BSs, broken BSs, and active UTs in defect cells. Table (c) shows alternative BSs for UTs in defect cell (Pattern A), which is used in Section IV.

Table I
SIMULATION CONDITIONS.

BS Position	on a circle with radius r_b and rotatio θ $r_b \sim U[0, 0.4]$ $\theta \sim U[0, 2\pi)$
UT Position	Uniform Distribution in Entire Cell
Defect Cells	$M_d = 3$
(BS,UT) Antenna Number	$(N_{b,m}, N_{u,m}) = \begin{cases} (14, 2) & L_m = 2 \\ (21, 3) & L_m = 3 \end{cases}$
Modulation	QPSK
SNR	$\text{SNR}_m = 10 \sim 30$ dB (default : 20dB)
Path Loss Exponent	$\alpha = 3.5$
Shadowing	Log Normal Distribution Standard Deviation $\sigma = 6$
Fading	i.i.d. Quasistatic Rayleigh

BS in each cell and one UT connects to each BS. Here, we deal with the cooperation of $M = 7$ cells (numbered by $m = 0, \cdots, M-1$) in downlink, and BS_m with $N_{b,m}$ antennas transmits its data $\{s_{m,\ell}(t);\ \ell = 0, \cdots, L_m - 1\}$ using transmit weight vector $\boldsymbol{w}_{b,m,\ell}$, and after passing through MIMO channel $H_{m,n} \in \mathbb{C}^{N_{u,n} \times N_{b,m}}$, UT_n with $N_{u,n}$ antennas receives it using receive weight vector $\boldsymbol{w}_{u,n,\ell}$.

In this study, two disaster patterns are considered: in pattern A, M_d BSs in the connected cell region are broken. In pattern B, M_d cells with broken BS are distributed among working cells. If the BS is broken, UT of that cell should connect to a BS in an adjacent cell, and the strength of link becomes weak. Our aim is to recover the communication quality by the cooperative work of BSs, and the performance improvement is evaluated not only in the cell edge but in the entire cell.

B. Design

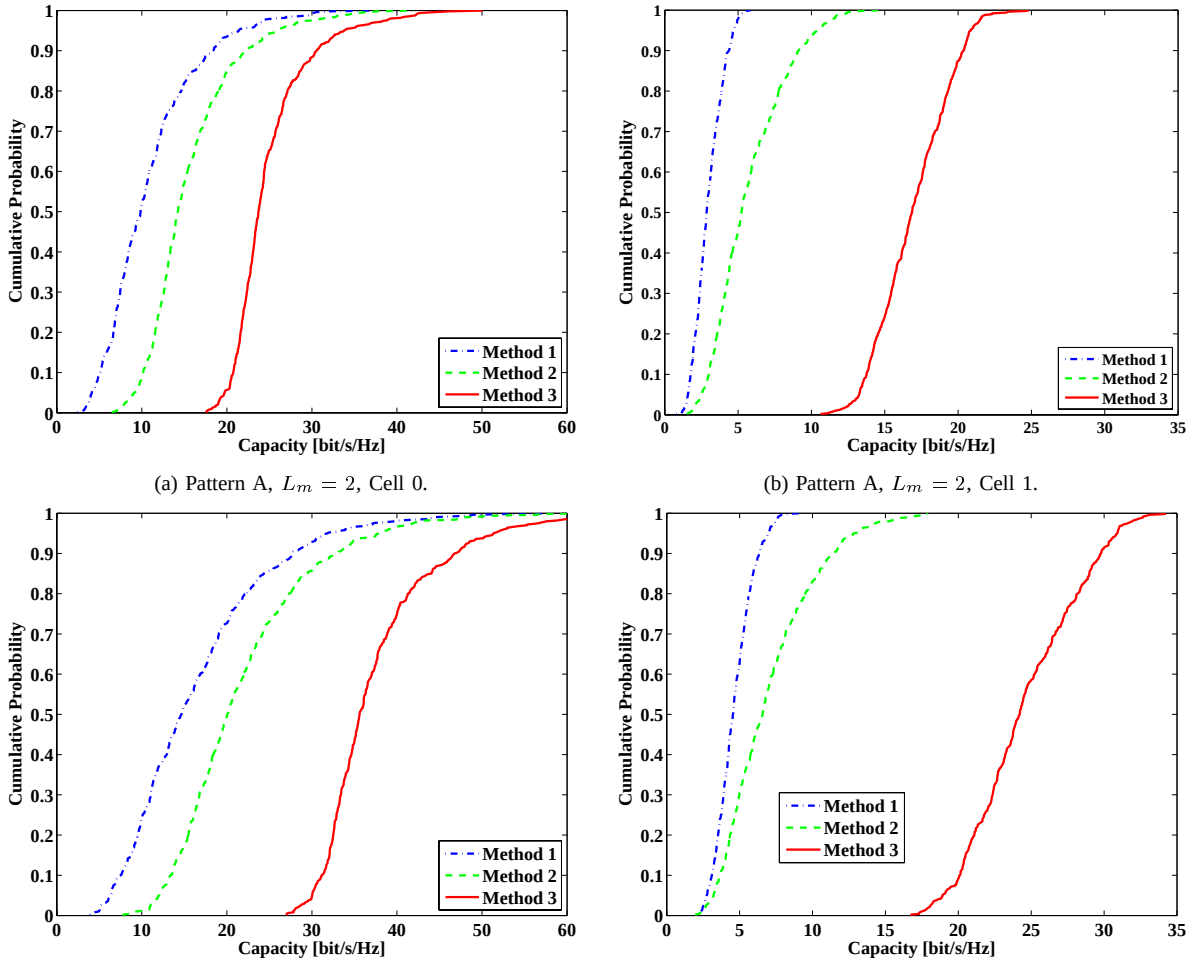
Three kinds of cooperative and noncooperative algorithms described here are same as previous studies except that they are multistream version, and selected for the main objective of this paper, CoMP effect evaluation under disaster situation. In this paper, CoMP scheme means the cooperative work of BSs utilizing the shared channel state information (CSI) and transmission data among BSs.

Method 1: This method does not consider the cooperation of BSs, and BS_m knows only its own channel $H_{m,m}$. First, the transmit weights are designed user by user like single user design, namely, $\{\boldsymbol{w}_{b,m,\ell}\}$ are designed by singular value decomposition (SVD) of $H_{m,m}$ (utilization from the largest to L_m-th largest singular value and related vectors). Then $\boldsymbol{w}_{u,m,\ell}$ is designed for the beamforming to minimize the sum of the power of all the signal except $s_{m,\ell}(t)$. In this method, the transmission interference mitigation to other users are not paid attention.

Method 2: This method considers the cooperation of BSs by sharing only CSI, and BS_m knows channel $\{H_{m,n};\ n = 0, \cdots, M-1\}$ including those of nontarget UTs. The receive weights are first designed by SVD (utilization from the largest to L_m-th largest singular value and related vectors). Then $\boldsymbol{w}_{b,m,\ell}$ is designed to steer nulls to all the undesired stream except the target weight $\boldsymbol{w}_{u,m,\ell}$ (zero forcing which consumes only one degree of freedom for the nulling to one stream). By this method, the transmit interference mitigation to other users could be achieved.

Method 3: This method considers the cooperation of BSs, and BS_m knows, in addition to CSI in Method 2, the data of all users $\{s_{m,\ell}(t);\ m = 0, \cdots, M-1, \ell = 0, \cdots, L_m - 1\}$ through backhaul link. By this condition, all the (working) BSs can construct a virtual array of $\sum_{n \in \mathcal{A}} N_{b,n}$ antennas, where $\mathcal{A}$ is the set of cell number with a working BS, which means the enhancement of desired link is possible avoiding the interference to undesired users utilizing sufficient degree of freedom. The transmit and receive weights are designed by block diagonalization [9].

 ISBN: 978-1-61208-203-5



(a) Pattern A, $L_m = 2$, Cell 0.

(b) Pattern A, $L_m = 2$, Cell 1.

(c) Pattern A, $L_m = 3$, Cell 0.

(d) Pattern A, $L_m = 3$, Cell 1.

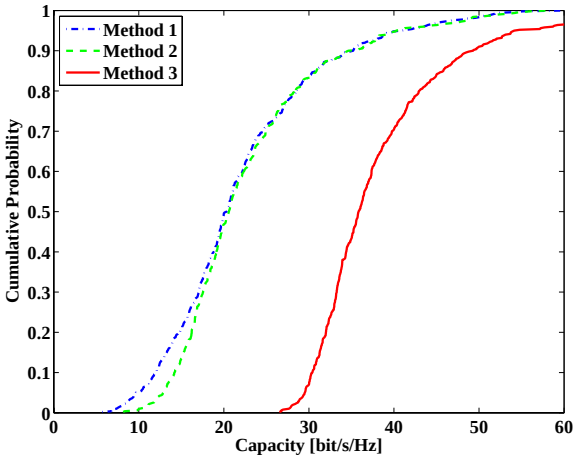
Figure 2. Distribution functions of sum capacity in two and three stream transmission for Patter A (SNR $= 20$dB).

IV. SIMULATIONS

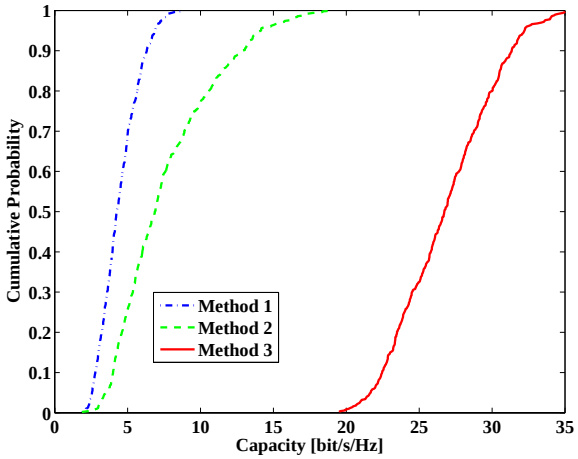
In this section, the performance improvement by cooperative communication is verified through computer simulations using algorithms in the previous section. The default simulation conditions are summed up in Table I. In [6], UT position is restricted to the cell edge, but it is removed here since the CoMP under disaster recovery is required to work for the user anywhere in the cell. To avoid the complexity of resource allocation, here we assume that BSs have the enough power to allow $P_{s,m} = 1$ for the transmission to each user even if it is connected to multiple UTs (though more practical evaluation including this problem is an important future work, our experiment is sufficient to measure the CoMP effect under the equal conditions for Method 1$\sim$3). The BSs have enough number of antennas for steering zero to all the users, but UTs are equipped with $N_{u,m} = 2$ or 3 antennas because of the limitation of the physical size.

Figure 2 (a)$\sim$(d) plot the distribution functions of capacity in pattern A for the system with $L_m = 2$ and $L_m = 3$ streams, respectively. In those subplots, (a), (c) and (b), (d) correspond to Cell 0 (BS is working) and Cell 1 (BS is broken). It is first verified that shapes of curves are not so much different between (a)-(c) and (b)-(d) though the actual total capacity is significantly increased in (c) and (d) by the use of larger number of eigenpaths. Between Cell 0 and Cell 1, we can find the gap of capacity: in Cell 1 which uses BSs of adjacent cells cannot achieve sufficient performance improvement even by using cooperation of Method 2 in which all BSs have CSI of all the users. This is because the conventional CoMP schemes are designed to mitigate the generation of interferences to nontarget users utilizing CSI, but our problem is rather in the weakness of the target signal. On the contrary, Method 3 which assumes the share of CSI and data of all the users by all the cooperative BSs achieves much higher improvement since the large size virtual array has also the effect of the enhancement of the transmitted signal. Another feature of the curve of Method 3 is that it

 ISBN: 978-1-61208-203-5



(e) Pattern B, $L_m = 3$, Cell 2.



(f) Pattern B, $L_m = 3$, Cell 1.

Figure 2. (Continued.) Distribution functions of sum capacity in two stream transmission for Patter B (SNR $= 20$dB).

has less steep gradient than that of others, which means the variance of the capacity becomes larger. This fact means that the quality is less stable, but the average capacity of Method 3 is more than 2.5 times of others, and it is much advantageous also in the point of outage capacity. Likewise, distribution functions of capacity in case of pattern B are given in Figure 2 (e) and (f). In subplot (e), characteristics of Cell 2 with working BS has a different behavior from (a) and (c) (curves of Cell 0 are similar to (a) and (c)): Cell 2 is located in the edge of the seven cells, and the influence of interferences is small, hence improvement by Method 2 is not anticipated. In Cell 1 with broken BS, the overall trend is not so much different from (c) and (d) except that the curve of Method 3 in (f) shifts to the right. This upgrade happens because in Pattern B, Cell 1 is surrounded by working cells which become the origin of source signal (in Pattern A, this is not consists), hence the cooperative transmission can invoke its advantage more effectively.

Figure 3 draws the relation between the input SNR and sum capacity for $L_m = 3$. Almost linear characteristic of curves of Method 2 and 3 means those algorithms well avoid the influence of interferences and synthesize the desired signal as strong as possible. On the other hand, without cooperation, such a linear improvement is not achievable, since the number of antennas in UTs is not enough to separate the signals from all the BSs. What is also remarkable is that the result of Method 2 (with CoMP) is worse than Method 1 (w/o CoMP) in low SNR region, what is not seen in the case of the conventional application [3]. The reason is considered as follows: The UT location is distributed over the entire cell and some of the origine of interference are destructed. Hence the noise becomes much dominant in the low SNR region, and Method 2 adopting ZF degrades due to high noise level [9].

The capacities other than $M_d = 3$ are given in Fig. 4.

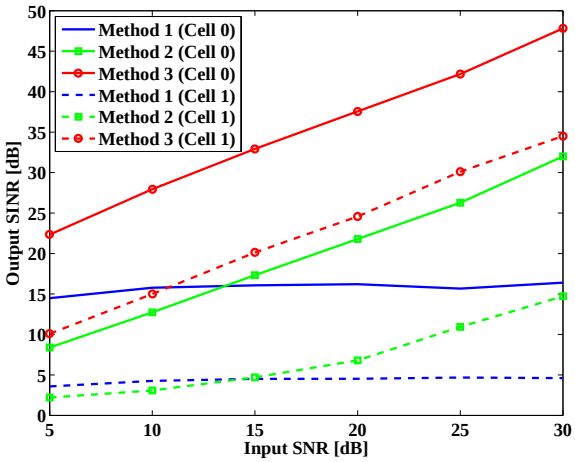
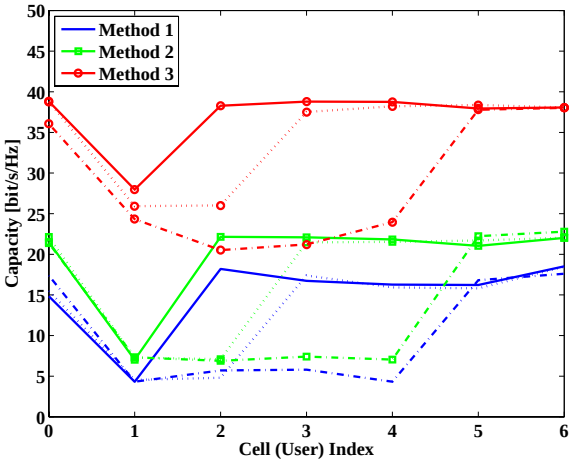


Figure 3. Input SNR versus sum capacity for $L_m = 3$.

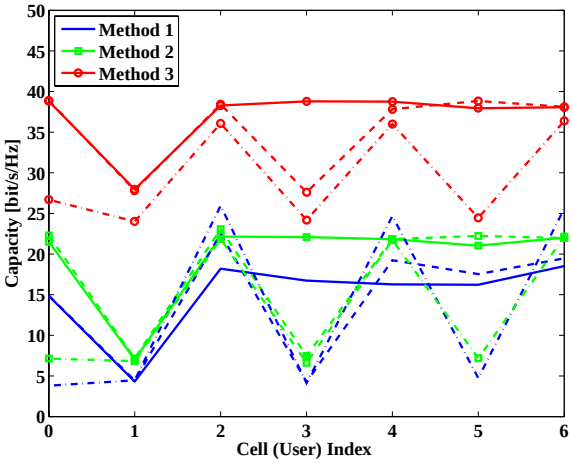
We can observe the performance degradation in defect cells and the recovery by CoMP scheme. In (a), as M_d becomes larger, the capacity difference against working cells is not well improved since there's few BSs which can be utilized for cooperation, while the gap of them improved in (b) where defect cell is well surrounded by the working cells, the origin of desired signal.

From those results, we can conclude that the CoMP scheme is effective for the disaster recovery also in multistream case. As the reason of the advantage of CoMP scheme, it is considered that enhancement of target signal is dominant to interference cancellation, hence the virtual array sharing CSI and data brings significant performance improvement, and its effectiveness compare to Method 2 is larger than in case of conventional base station cooperation [3]. On the contrary, the effect of CoMP sharing only CSI

 ISBN: 978-1-61208-203-5



(a) Pattern A.



(b) Pattern B.

Figure 4. User index versus sum capacity for $M_d = 1$ (solid line), $M_d = 2$ (broken line), and $M_d = 4$ (dashed line).

is not sufficient in this application. What we can learn is that it is desirable to connect BSs around disaster risk area through backhaul link, which reinforces the result of [6]. Though the results are for downlink phase, they provide enough materials to infer the advatage of CoMP also in case of uplink scenario.

V. Conclusion and Future Works

This paper has evaluated the multistream cooperative communication scheme for the disaster recovery in MIMO cellular system where BSs in some cell are broken. We have considered typical two types of disaster-suffered cell patterns, and three kinds of cooperative and noncooperative algorithms, and evaluated the performance improvement by cooperative transmission in downlink through computer simulations under various conditions. The results show that the concept of CoMP scheme utilizing survived infrastructure is effective also for performance recovery in disaster area.

The future work is the relay-aided processing: in this case the relay station may be a portable type (e.g., mounted on a vehicle) and to keep the fairness of the user connection, mobile relay might be more suitable. In addition, the resource allocation putting importance on the disaster area becomes an important theme of the study.

Acknowledgement

Authors will express thanks to Dr. Teruya Fujii, Soft Bank Telecom, Tokyo, Japan. This work was partially performed under research contract on cooperative base station system for the Ministry of Internal Affairs and Communications (MIC) of Japan.

References

[1] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa, and M. Tanno, “Coordinated multipoint transmission/reception techniques for LTE-advanced,” *IEEE Wireless Commun.*, vol. 17, no. 3, pp. 26-34, June 2010.

[2] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, “ Coordinated multipoint: Concepts, performance, and field trial results,” *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102-111, Feb. 2011.

[3] T. Taniguchi, Y. Karasawa, and N. Nakajima, “Performance amalysis of base station cooperation in multiantenna cellular system,” *IEICE Trans. Commun.*, vol. E94-A, no. 11, pp. 2254-2262, Nov. 2011.

[4] E. Bjornson, N. Jalden, M. Bengtsson, and B. Ottersten, “Optimality properties, distributed strategies, and measurement-based evaluation of coordinated multicell OFDMA transmission,” *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6086-6101, Sept. 2011.

[5] Y. Akaiwa, “An adaptive base station cooperated cellular system and its theoretical performance analysis,” *Proc. 2011 IEEE 73rd Veh. Technol. Conf. (VTC2011-Spring)*, Budapest, Hungary, May 2011.

[6] T. Taniguchi, Y. Karasawa, and N. Nakajima, “Base Station Cooperation in Multiantenna Cellular System with Defect Cells,” *Proc. 2011 7-th Loughborough Antennas & Propagat. Conf.*, Loughborough, U.K., Nov. 2011.

[7] D. Camara, N. Frangiadakis, F. Filali, A. Loureiro, and N. Roussopoulos, “Virtual access points for disaster scenarios,” *Proc. IEEE 2009 Wireless Commun. Networking Conf. (WCNC 2009)*, Budapest, Hungary, Apr. 2009.

[8] F. R. Yu, J. Zhang, H. Tang, H. C. B. Chan, and V. C. M. Leung, “Enhancing interoperability in heterogeneous mobile wireless networks for disaster response,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 5, pp. 2424-2433, May 2009.

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

 ISBN: 978-1-61208-203-5

Wireless Module for Data Collection

Alexey Lagunov
Department Computer Science and Electronic Devices
North (Arctic) Federal University
named after M.V. Lomonosov
Archangelsk, Russian Federation
a.lagunov@narfu.ru

Dmitry Fedin
Department Computer Science and Electronic Devices
North (Arctic) Federal University
named after M.V. Lomonosov
Archangelsk, Russian Federation
d.fedin@narfu.ru

***Abstract*—At the present stage of development various measuring and executive devices (meters, sensors, actuators, etc.) are widely used and these devices are increasingly equipped with one of the standard interfaces such as USART (RS232), SPI, 1 - Wire, I2C, etc. The fact is given opportunity to project and make the extensive distributed multipurpose automatic control systems with complicated control algorithms and to automate the process of picking up and centralized data processing. But, solving this problem requires the use of small low-cost embedded microcontroller systems as buffer management, as well as GSM and Wi-Fi modules for the organization of a communication channel with the data collection and processing center. The use of microcontrollers as the buffer allows for some level of required operations to produce and primary processing of data and prepared directly at the control point. We designed the module - embedded controller of a remote data acquisition and control devices/equipment.**

Keywords-wireless; module; data; collection

I. INTRODUCTION

Nowadays, computer-aided data acquisition systems are accessible way to get the experimental date, and it is connected, first of all, with a wide spread of personal computers. Data acquisition systems are used for scientific research, production process management, industrial monitoring, medicine, meteorology, astronautics and others fields of human activities. Computer-aided data acquisition suggests the new quality of data, which is impossible to get by other way – it is result of measurement abundance statistical treatment an in the digital form; opportunity of registration accidentally appearing event with unattainable earlier resolution in time and amplitude; fast processes registration. Due to the quick reduction in the data acquisition systems price in comparison of human effort cost it is found the usage in the area, in which uses the hand-operated data registration: greenhouse, elevator, meteorological station, the process of acceptance-and-transfer and certification production test, storehouse, industrial cold-storage plant, boiler room, science experiment automation, etc.

As an analogue of our module considers the following devices:

- TWCT20;
- AirLink GL6100.

AirLink GL6100 [5] - RS232 wireless gateway using the communication channels GSM \ GPRS (850/900/1800/1900 MHz frequency bands). Allows working with interfaces UART (one piece). Supports hot plugging controlled devices. It is economy (Rated current 3 mA, max - 400 mA). Use as a powerful and cost-effective CPU ARM946 with a frequency of 104 MHz allows to execute custom scripts in languages C \ C + + and Lua.

The positive aspects:

- Economy;
- supports all frequency ranges for the GSM \ GPRS channels used in Russia;
- Powerful efficient processor;
- Compatible with OS Linux, Windows.

The negative aspects:

- Few supported Interfaces;
- Low maximum power GSM \ GPRS transmitter;
- No possibility change the firmware remotely.

TWCT20 [3] – Wireless terminal of remote control devices \ equipment. For connection using GSM \ GPRS networks (850/900/1800/1900 MHz). Lets you take the signals from analog sensors, digital input lines, to work with RS232 interface and manage the digital outputs (250 V 7A). Supported protocols - SMS (text mode), HTTP, SMTP. Configuration is done through an internal Web-based interface.

The positive aspects:

- High speed of data exchange (16-48 KB / s);
- Working with the two analog signal sources;
- Commutation of power load (up to 7A).

The negative aspects:

- - No support for user scripts;
- - Low operating temperature range (-20 - 55);
- - High operating voltage (24 V).

Our module - embedded controller of a remote data acquisition and control devices / equipment. For communication using GSM \ GPRS networks (850/900/1800/1900 MHz), Wi-Fi, ISM (2.4 GHz), Ethernet. You can take the signals from analog sensors, sensors that use the interfaces RS232 (2 pieces), SPI, CAN (1 piece), I2C. When you work uses custom scripts loaded from the server without the user. Supports the SMS, HTTP, TCP \ IP, SMTP, FTP, FAX.

The positive aspects:

 ISBN: 978-1-61208-203-5

- Multiprotocol;
- High data rate (up to 115,200 baud);
- Wide range of operating voltages (5-24 V);
- Efficiency (rated current of 40 mA);
- A standby power supply (up to 5-day battery life);
- 2 RS232 interface;
- Modularity;
- High sensitivity GSM \ GPRS receiver;
- Possibility to control an external power load;
- Wide working temperature range (-40 to +85);
- Possibility of caching data.

The negative aspects:

- Low CPU performance;
- No opportunity to work as a gateway (only via CSD for GSM).

Section "Data acquisition systems" is devoted to reviewing the available data collection systems. In the "Wireless module ", we consider the design features of the wireless module for data collection. On the basis of the developed theory, we lead a experimental research that is given in Section "Testing".

II. DATA ACQUISITION SYSTEMS

Data acquisition systems may be used in the real-time regime, for instance, for monitoring different process, emergency conditions identification in technological systems, for management, and also for data archiving, when they are separated from the processing procedures for the collection during a time interval. In the real-time systems, current data save in circular buffer, while older data displace the new date, during the current time. Information tank of greater capacity is used in archive systems, and data is processed after data acquisition completion.

Archive data acquisition systems (loggers, recorders) may be self-contained unit, constructed on base of microcontroller, for instance, airborne recorder, electronic counter of heat or electric power, portable electrocardiograph). Data collected by loggers is transferred to processing to computer with help, for instance, of a USB flash memory or through serial port. Archive data acquisition systems (loggers, recorders) may be self-contained unit, constructed on base of microcontroller, for instance, airborne recorder, electronic counter of heat or electric power, portable electrocardiograph). Data collected by loggers is transferred to processing to computer with help, for instance, USB flash memory or through serial port.

A data acquisition systems constructed on the basis of computer allow collecting and processing data at the same place and often with the help of the same software. This is the most widespread version of such systems performance. The wide possibilities for collecting and processing data are presented by MatLab, LabView, MS Excel.

Systems with parallel bus, including PCI-cards, are used for fast processes registration (with required sampling rate more than 1 MHz). Computer boards have limited number of input, defined by constructional specification, and require external terminal block for connection of signal source, which is inconvenient at mounting system.

External devices connecting to computer with such ports as COM, USB or Ethernet are more convenient for slow process registration. External devices are different by less noise, while the card inserted into a PC, are influenced by interference from the computer's digital circuits.

Data acquisition systems may be distributed, when devices input block is separated on Data acquisition object territorially, and receiving data collect to one storage and data transfer with help of network technologies (Ethernet, Modbus, Profibus, DeviceNet, CANopen, DCON and others, wireless network Bluetooth, Wi-Fi, ZigBee, Internet technology, intranet). Distributed data acquisition allows particular uncounted increasing number of inputs channel; however it is restricted by network date rate. Data acquisition systems may be distributed, when devices input block is separated on Data acquisition object territorially, and receiving data collect to one storage and data transfer with help of network technologies (Ethernet, Modbus, Profibus, DeviceNet, CANopen, DCON and others, wireless network Bluetooth, Wi-Fi, ZigBee, internet technology, intranet). Distributed data acquisition systems allow particular uncounted increasing number of inputs channel; however it is restricted by network date rate.

Data acquisition systems input may be universal (current, inductive and potential) and specialized (for instance, for thermocouple, thermoelement resistance or tensormeter). System with specialized input is economically effective to user. Universal input uses together with measurement converter of physical value to current and voltage. There are system with hybrid input, for instance, when one input received the thermocouple signal, other input – tensormeter signal, third - thermoelement resistance and etc.)

Inputs may be differential, single or digital. Differential input allow more effectively suppress internal noise inducing to cable transmitted signal from detecting device to input module. Voltage in the range ±(0...5), ±(0...10) V or current in the range of 0..20, 4...20 mA is used the data transmission. Voltage signals is worked out by voltage source and have the high noise immunity to capacitive pickup; current signal is worked out by current source and stable to inductive pickup. Digital inputs receive logical signals (“0” or “1”) arrived from limit switch, intruder or fire alarm sensor, electromagnetic relay, voltage presence sensor and etc [1].

The major settings of data acquisition systems are channels number, inaccuracy, dynamical inaccuracy, establishing time or pass band, resolving power, effective digit count, sampling rate, galvanic input and interface isolation availability, availability of defense of careless usage, overload and overheating.

Generally data acquisition systems have 4, 8, 16, 32, 64 ... input, inquired by turn or simultaneously. System with simultaneous inquiry is consist of identical channel, which is done the analog-digital conversion of input value at the same time for all channels. Such system is uncommon due to expensive cost. Generally input inquiry performs in turn with help of commutator. Therefore different channels data is shift to time on the delay equal to relations of inquiry time to channels quantity [2].

 ISBN: 978-1-61208-203-5

III. WIRELESS MODULE

There is a few variant of data acquisition systems in the radio technical monitoring center North (Arctic) federal university named after M.V. Lomonosov.

As a mentioned above at the present stage of development various measuring and executive devices (meters, sensors, actuators, etc.) are widely used and these devices are increasingly equipped with one of the standard interfaces such as USART (RS232), SPI, 1 - Wire, I2C, and etc. The fact is given opportunity to project and make the extensive distributed multipurpose automatic control system with complicated control algorithms and to automate the process of picking up and centralized data processing. But to solve this problem requires the use of small low-cost embedded microcontroller systems as buffer management, as well as GSM and Wi-Fi modules for the organization of a communication channel with the data collection and processing center. The use of microcontrollers as the buffer allows for some level of required operations to produce and primary processing of data and prepared directly at the control point. In our case, a system based on high-quality GSM modules manufactured with built-in powerful Telit ARM microcontroller and the virtual machine interpreter a powerful Python. In addition to a sufficiently powerful processing core, this module is implemented by hardware support 2 USART interfaces and one SPI, as well as a number of universal input-output ports for user purposes. The solution allows to have available a flexible system with a powerful and versatile scripting language. But the use of GSM channel is mainly justified when applied to mobile and very distant objects. When used on most real-world objects and within the city limits in most cases it is more efficient use of wireless communication channel based on the Wi-Fi technology as in this case is easier and cheaper to hold a LAN to a monitored object and do not pay for the services of mobile operators GSM / GPRS bands. In the result of led researching-constructive works the system presented on follows figure is worked out.

The general system operation principle is displayed in the picture in general (Fig. 1). Seen from the figure that the system can adapt itself to a wide range of tasks, changing only the software on the server, data collection terminals to client management, and scripts embedded microcontroller, without replacing the hardware system. In this case, the user can select the most appropriate for him the connection channel.

The implemented the center module is constructed using GSM processor Telit GL865-Dual and Wi-Fi module SM2144N2 (Fig. 2).

As an analogue of our module considers the following devices: TWCT20 [3], Cinterion DSB75 [4], AirLink GL6100 [5]. Considered analogues have similar characteristics, but have, in our view, a number of disadvantages: impossibility reprogram and change the settings of the driver remotely. Usually the remote parameters change function is supported only; a supported interface type (RS232).

We have tried to correct the disadvantages of the data. Our module is:

- As part of the total data collection system can not only change the remote settings of the device, but also the change of the remote driver to work with different types of devices. Changing the driver can happen automatically, without human intervention;
- It supports not only the interface RS232, but also widespread in industry interface CAN;
- There is an additional +5 V power supply output for powering external sensors or, if necessary, power interfaces, plug-ins.

IV. TESTING

As the unit tests were analyzed levels of the signal module and 3G modem (Huawei E173) in the frequencies of the network GSM. The measurements were performed using the same SIM-card and the investigation object were placed in the same point-space by rotation. Measuring levels the signals was carried out in automatic mode with the help of the same the software (the script in Python, runs with the object at the level of AT commands) in order to exclude the human factor. Measurements were carried out within 10 minutes with an interval of 10 seconds. The results are shown in the graph (Fig.3).

As seen from the results, our module has a higher level signal from a base station, which allows for higher rate data transfer.

V. CONCLUSIONS

System can adapt itself to a wide range of tasks, changing only the software on the server, data collection terminals to client management, and scripts embedded microcontroller, without replacing the hardware system. In this case, the user can select the most appropriate for him the connection channel.

[1] Bailo C., Alderson G., Yen J. Re-quirements for Open, Modular Architecture Controllers for Applications in the Automotive Industry//www.isa.org.

[2] Fischer H., et al. The COMPASS Data Acquisition System//IEEE Trans. Nuclear Science, 2002, v. 49, .2, April, pp. 443–447.

[3] http://russian.gsm-gprs-modem.com/china-programmable_gateways_and_gsm_gprs_modems_for_m2m_sierra_wireless_airlink_gl6100_and_gl6110-130233.html

[4] http://www.sectron.eu/products/1-cinterion-wireless-modules/186-development-kit/919-cinterion-dsb75-development-kit.html

[5] http://www.connect-gsm.ru/oborudovanie/teltonika/61-twct20-programmiruemyj-gsm-kontroller

 ISBN: 978-1-61208-203-5

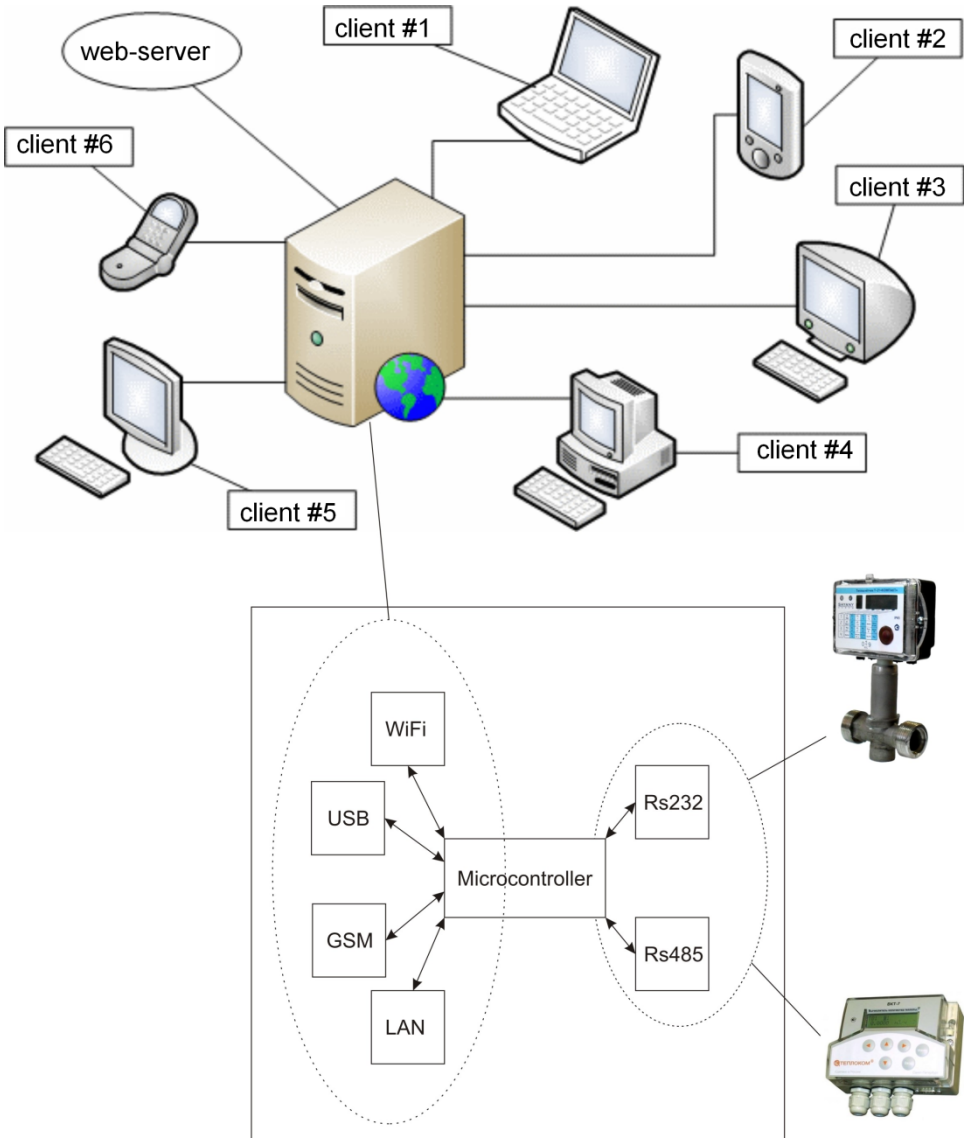


Figure 1. The general system operation principle.

Figure 2. Module sample

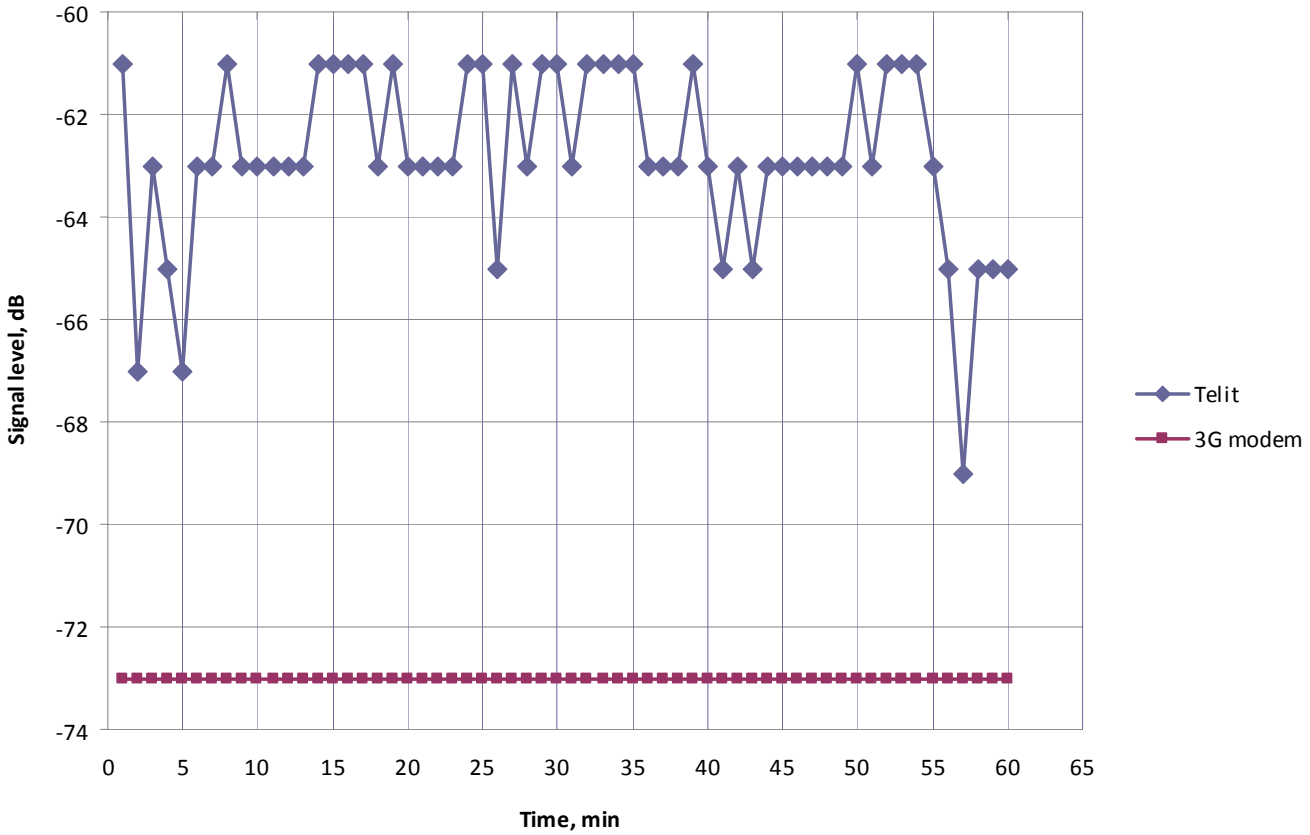


Figure 3. Test results

 ISBN: 978-1-61208-203-5

Communication aspect in ICT for Freight Transport System

Pushpendra Kumar, Belkacem Ould Bouamama
LAGIS UMR CNRS8219
Ecole Polytechnique Universitaire de Lille
Villeneuve d'Ascq, France
jahan.pushp@gmail.com
Belkacem.Ouldbouamama@polytech-lille.fr

Haffaf Hafid
Faculty of Science, Computer Science Department
Oran University
El M'Naouar, Oran, Algeria
Haffaf.hafid@univ-oran.dz

***Abstract*—In the last decade, Information and Communication Technology (ICT) proved to be a milestone in the field of freight transport with the appearance of tracking and tracing devices. Different ICTs viz. Radio Frequency Identification (RFID), wireless sensor nodes and localization systems play vital role to improve the performance of the freight transport system by saving energy consumption, reducing the service cost and increasing the cargo throughput. To achieve these requirements, the development of reliable heterogeneous communication system among all communicating objects becomes a paramount objective. In this paper, we describe different kind of existing communication technologies in Intelligent Transport System (ITS) domain. Also, we propose an intelligent infrastructure integrated with ICTs for the operations of Intelligent Automated Vehicle (IAV) in confined space like container terminal.**

Keywords-Freight transport system; Intelligent transport system; Information and communication technology; Intelligent automated vehicle; Port conatiner terminal.

I. INTRODUCTION

The convergence of conveying, storing and manipulating data has led to many new and exciting developments in the Intelligent Communication Technology (ICT). An Intelligent Transport System (ITS) is advanced application of ICT which aims to provide innovative and efficient services relating to different modes of transport. The optimal transport strategy based on real-time information, as well as assessments of the transport system demand and supply should be given more importance in order to improve the productivity of the system. In the information society, the ICTs are developing towards an infrastructure that will enable new kinds of practices also affecting the transport system, for example trucks integrated with advanced devices that link the driver cab to the haulage operators' systems.

Intelligent Automated Vehicles (IAVs) are cornerstone of the future ITS and can be seen as the logical transition of mobile robotics to the scale of vehicles in urban transport or in container terminals. In the context of urban transport, by enabling vehicles to communicate with each other (Vehicle-to-Vehicle or V2V) as well as with base stations (Vehicle-to-Infrastructure or V2I) via wireless communication networks can contribute to safer and more efficient roads by providing timely information to drivers and concerned authorities.

The problem in freight transport systems is to obtain information in ITS over Mobile Relay Network to facilitate the necessary information access for drivers on the road. The proposed network solution consists of all the RFID enabled mobile nodes on road such as RFID tag and reader on the traffic light pole, and Wireless Sensor Network (WSN). This enables cars to be aware of their position and of the vessel they transport. In our case, we are limited with a small platform such as seaport confined space where cars are IAV nodes, and further works aim to generalize to more complex infrastructure.

This work is performed in part of European project Weastflows [24]. The paper is organised as: In Section II, the role of ICT in sustainable transport system is described; then, in Section III, various ICTs are explained with their applications in freight transport system. In Section IV, a case of communication in container terminal is discussed. Concluding remarks are lead to the last section.

II. ROLE OF ICT IN SUSTAINABLE FREIGHT TRANSPORT

The key role of ICT for enabling sustainable freight transport is in establishing cooperation among logistics companies and various actors of freight transport system by enabling the real time flow of information. ICT helps to build trust among the various actors of the freight transport system by encouraging them to share information for achieving optimum transport strategies. Another important contribution of ICT towards sustainable freight transport system is its ability to support intermodal freight transport. Various practical applications of ICT in freight transport include vehicle tracking, monitoring and control, vehicle to vehicle communication, vehicle to infrastructure communication, security and safety purposes.

Ollo-López and Aramendía-Muneta [4] examined the impact of ICT on competitiveness, innovation and environment, and found that use of ICT seems to favor these issues. Several applications can be cited thanks to ICT development in ITS: Tracking and tracing, localization, Monitoring and control, dynamic scheduling, traffic flow (optimization**),** weather and congestion information, pollution control, safety and security. For example, a trailer could be automatically identified, given permission to enter

 ISBN: 978-1-61208-203-5

a container yard and instructed where to drop its load. In [20], a framework of a devoted highway freight transport platform in China is described. Such platform can provide the availability of drivers to transport companies by destination requests, generating route plans, and returning the calculated plans to the users. Furthermore, the information is processed through a Geographic Information System (GIS) [19] capable to provide accurate and real-time weather information on a specific area.

There are recent trends towards Vehicular ad hoc Network (VANET) as they can leverage mobile nodes to bridge the gap between information isolated islands. It is a flexible and low-cost extension of wired infrastructure networks. With its ubiquitous feature, VANET is attracting intensive interests in many application areas. Nowadays, the spread of unmetered high-speed connections of internet allows greater flexibility and can be realized in working location. The Internet can be accessed almost anywhere by numerous means, including through mobile internet devices, mobile phones, data cards and cellular routers.

Figure 1 shows the flow of information in a freight transport network. All the information regarding the transport network is communicated to a central data base, which is updated with the current situation of the transport network. This information includes location of the cargo, congestion, incidents and vehicle breakdown. This information is useful for the different actors of the transport system, so that they can make the better business strategies according to the current situation of the transport network. This information flow allows real time control of the network to achieve the goal of sustainable transport system.

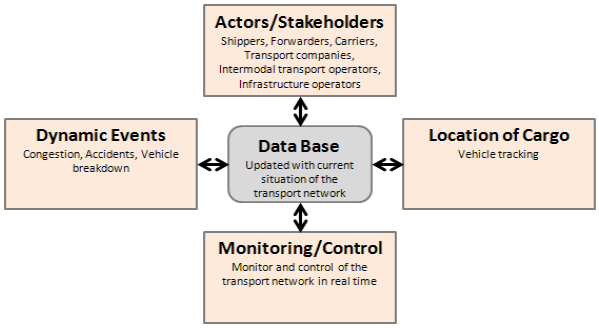


Figure 1. Information Flow in Freight Transport Network

However, some limitations have to be taken into account when choosing some plate-form. First of all, the quality of information delivery due to limited range in wireless protocols: Most conventional ITS technologies can only detect the vehicle in a fixed position. The second impediment is the lack of security data exchange, especially in wireless case. Coronado et al. [23] proposed secure service architecture for logistics covering road haulage, infrastructure and port operations. The various problems can occur with the implementation of technology including time delay to provide reliable information in critical situations, some network access are free and other not, lack of standardization especially when interoperability is needed, reliability, scalability (a huge number of vehicle), implementation cost and network security.

III. ICT TOOLS FOR FREIGT TRANSPORT SYSTEM

In this section, we discuss about the various technologies used in freight transport system. Vehicle telematics system which involves telecommunication aspects like GPS, or more generally web-based information system can be used in trailer tracking and on-board navigation. Also, RFID is a type of automatic identification technology, has been increasingly used with a great success since the 1990s [2] and RFID-RTLS (Real time locating system) is well suited for container tracking. Infrastructure-based wireless communication has experienced a huge diversification of radio access technologies while experiencing a steady increase of capacity. Table I shows different technologies used in freight transport.

TABLE I. ICT TOOLS IN FREIGHT TRANSPORT

Communica-tion Mode	**Most used tools**	**Applications**
Sensors	-Intrusive -non intrusive -embedded	Geo-fencing, Parameters evaluation: speed, distance, Safety guidance
Diffusion information	Panels or bulletin board electronic panels High speaker	Highway example
Localization	GPS, Galileo, Hertzien network, Vehicle Telematics	Vehicle tracking and tracing, Navigation, Geofencing
Tracking Identification and tracing	RFID, GSM 3+, RTLS, Gyrometer, Inertial center	Container tracking, Individual product tracking
Cooperative systems	GSM UMTS	Communication (V2V, V2I, IAVs)
Dedicated Transmission	GPRS (General Packet Radio Service) Radio Electric specific transmission	Confined space, Automatic toll collection, Safety Warning

Beyond third generation (B3G), wireless networks comprise of highly ubiquitous and fast mobile broadband access technologies such as WLAN (Wireless Local Area Network), HSPA (High Speed Packet Access), or Mobile WiMax. Sensor nodes (installed on each vehicle) can collect information in order to organize the traffic, especially at intersection where we have not traffic light. Sensors should be installed on both roadside and intersections. Then, the embedded unit can send vehicle parameters (location, speed, direction, etc.) using ZigBee (IEEE 802.15.4) to roadside units. The challenges today are: how could be developed new solutions and integrate them in the existing

 ISBN: 978-1-61208-203-5

infrastructure communication and how can heterogeneous systems involving different technologies exchange information through the network, independently from the machines or devices they use i.e. interoperability.

Example of interoperable system communication is the it839/u-it839. Korea's "it839"3 project in 2004 has been one of the first national future Internet initiatives. It aims at funding research in eight different communication services (WiBro, DMB, home networks, Telematics, RFID-based, WCDMA, terrestrial DTV, and Internet telephony), three future network infrastructures (Broadband converged networks (BcN), soft infraware, the IPv6 architecture) and nine hardware-related businesses. From the point of view of services, the Location- and Context-Based Services (LCBS) are a new class of services with a high potential in the near future in freight transport. The key aspects of LCBS are their inherent relations between location coordinates or activity-contexts and applications.

Typical example classes for LCBS are:

• Intelligent navigation support for mobile users.

• Geography-dependent information systems and geographic multicast communications.

Configuration of ad hoc infrastructures for location or context dependent applications in case of rescue/emergency situations, spatially and time-limited operations requires a decentralized provision of location information within the terminal equipment through specific sensors (GPS, active badges, user input, etc.). The multi-network access of mobile terminal equipment (WLAN, GPRS, UMTS and Differential GPS) with horizontal and vertical handover capability can be used to improve the accuracy of the receiver by adding a local reference station to augment the information available from the satellites. Another technology RFID is mostly used in retailing and manufacturing environment but recently, it has been used in tracking vehicle as well, because using high frequency transmission enables readers to get information from great distance. The RFID chip typically is capable of carrying 2000 bytes of data or less. Table II summarizes different communications modes in RFID.

TABLE II. DIFFERENT RFID COMMUNICATION MODES

RFID Mode	Frequency /time	Characteristic	Application
Passive Time has non limited lifespan Tag energized by the reader	Low frequency	Least interference from metal, liquid	Access control (airport)
	High frequency	Tracking items everyday	Animal tracking
	Ultra high frequency	3.5 m and up Lowest price (no battery)	Pallet or case tracking (merchandise)
Active 200 m read-range 3-10 years tag life	Beacon	Presence or absence of items	Boolean applications or detection
	Real time location	The battery periodicaly transmits its signal	Find objects or people in real time

To resolve the problem of the large transmission overhead when the RFID tag information is transmitted in IEEE 802.11 wireless LANs, we can consider the frame aggregation method being discussed in the IEEE 802.11n Task Group. The frame aggregation method has two techniques called MSDU (MAC Service Data Unit) aggregation and MPDU (MAC Protocol Data Unit) aggregation.

For efficiently transmitting in real time RFID tag information in IEEE 802.11 wireless LANs, we need to combine the multipolling method [3], which enhances the PCF protocol using the connectivity information. The EPCglobal Network [14] (developed by the M.I.T) is the Auto ID international center's specification and specifies major aspects of operation of networked RFID system. The EPCglobal network architecture (middleware plate-form) can enable readers to identify and monitor RFID products and then, access crucial database to query cargos information shared through Internet. Furthermore, the system is regulated by international standard such as ISO 6346:1995 and ISO 17363:2007. In the traffic information management, RFID can be used to collect information about traffic jams. For instance, officials can track the travel time of cars on specific motorways, analyses that information and then distribute reports about average commuting times to drivers, helping them decide which route to take. There are two kinds of vehicle tracking: automatic vehicle tracking, when the vehicle transmits its location regularly (within time interval) and events activated tracking system when the tracking system is activated in reaction to some event [15].

Among tools for positioning problem or geo-localization, we can cite GPS (or DGPS), GIS and RTLS (Real Time Localization System). Also, optical character recognition (OCR) and biometric based technology are used for vehicle or product identification. A great number of physical embedded sensors exist to enhance safety guidance to driver (speed, distance, sound sensor etc.) and finally many kind of interface or electronic devices allow the driver to access to the infrastructure networks (computers, PDA, iPad, smart phone, embedded camera, digital billboard etc.). Some authors propose a Dedicated Short Range Communication (DSRC) such as the ETC system used to collect highway tolls. This promising wireless technology is capable of handling the information requirements associated to road transport. It could facilitate high-speed transmission of large volumes of data between a vehicle and equipment installed alongside the road.

IV. A CASE OF COMMUNICATION IN PORT CONTAINER TERMINAL

In the recent years, advanced ICTs and IAVs have been identified as possible candidates to implement Automated Container Terminals (ACT) to improve the performance of seaport management. New breakthrough of optimizing the

 ISBN: 978-1-61208-203-5

port logistics are achieved for smart container terminals. In an ACT, three types of handling facilities are usually used: Automated Yard Cranes (AYCs), IAVs and Quay Cranes (QCs). IAVs are autonomous vehicles used for horizontal transportation and are powered by batteries, which are automatically charged at a charging station on the terminal apron. To be aware of its localisation at any time within the confined deployment space, an IAV node can include (but not necessarily) a GPS circuitry. RTLS is more suitable in this case which provides the exact position of a node to which the RTLS tag is attached at each time. At least three RTLS readers should receive signals from these tags to calculate the position by means of an engine which collects continuously the signals. In our case, the tags could be attached to AYC to play the role of readers, or "Anchor nodes". Reduced Signals Strength Indicator (RSSI), Time of Arrival (ToA) and Angle of Arrival (AoA) give approximate location of the "tracked" nodes by using triangulation computation methods [10]. The second problem, in which RFID may play a major role, is the identification of an object tagged by a transponder. This RFID localization solution is based on radar principles [18], but it couldn't go further than 200m.

Now, we discuss the technologies to deal with the communications between IAV-to-IAV, IAV-to-CBS (Central Base Station) and IAV-to-RFID. These can be applied in freight transport infrastructure communication. The European project CVIS (Cooperative Vehicle Infrastructure Systems) has been investigating the capability to link vehicles to the roadside infrastructure through seamless communications channels. We distinguish two types of links in IAV-2-IAV and IAV-2-CBS communications: low rate wireless link to exchange data that do not require high transfer speed such as periodic and cooperation/coordination based data and high rate wireless link to exchange data of important volume such as data to be stored in data base or multimedia data, that are known as on demand data.

Because of possible interferences, IAV nodes could be equipped with a second transceiver such as GSM/GPRS technology to allow direct communications without passing by CBS nodes or another intermediate IAV nodes. Powerful meshing technical equipment such as Meshlium product proposed by Libelium^{TM} company [25] exists today in the market. This equipment allows interconnecting several wireless technologies at the same time (Wifi, Bluetooth, Zigbee, GSM/GPRS) and wired Internet. GPRS [14] has been introduced specifically for packet communication in GSM, it is described as 2.5G (between 2 and 3 generation) and recently enhanced by EDGE (Enhanced Data rates for global evolution). It is, thus, much more adequate for the application case under consideration. Contrary to voice/data communication in GSM (where an FD channel is reserved irrespective of the traffic intensity in the uplink/downlink direction) all users of a cell share the bandwidth.

For identification problem, we use RFID system in which typically a reader is placed on IAV and a tag attached to the container [8]. RFID reader and tag communicate wirelessly using antennas (see Figure 2). OCR (Optical Character Recognition) based system is also used for vehicle number plate's recognition. When a vehicle arrives at the terminal gate the OCR system captures its plate number, so that the control system records it together with the time and date, and verifies if the vehicle is allowed to get inside the area.

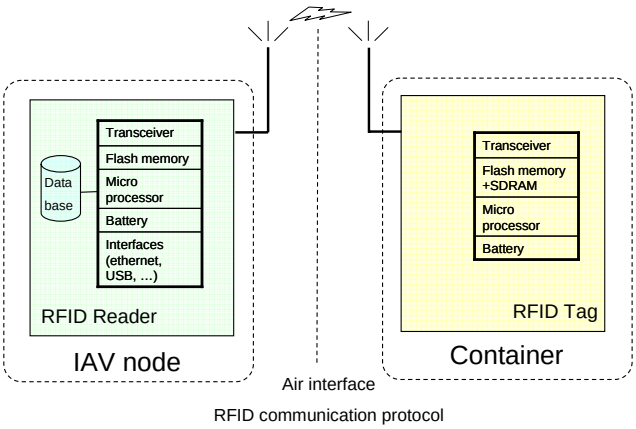


Figure 2. IAV- RFID Communication

With regard to the communication architecture, we see that there are many standard solutions related to RFID technology exists in practice [5][6]. Choosing the right RFID tag hardware able to ensure effective and efficient communication is critical to the overall solution. The choices depend on various factors like amount of container information to be stored on the tag, distance requirements, and confined space conditions. This communication takes place between each IAV and the corresponding container in loading/discharging operations using RFID. Real-time locating system can also be used for local positioning; it allows tracking and identifying the location of objects in real time. ISO/IEC 24730-5:2010 defines an air interface protocol which utilizes chirp spread spectrum (CSS) at frequencies from 2.4 GHz to 2.483 GHz.

Cellular systems such as UMTS are also good candidate for IAV-2-CBS communications because it accounts for protocol design mobility and data rates can reach up to 21 Mbps with high speed downlink packet (HSPDA). The imminent 4G technologies include Mobile WiMAX, Long term evolution (LTE) and HSPA+. The estimated coverage range is about 3 Km with a star network topology, LTE has a theoretical top download speed of 300Mbps and an Upload Speed of 75Mbps. Another candidate for IAV-2-CBS is Digital Media Broadcasting (DMB) which includes digital audio /video Broadcasting (DAB/DVB) however it only relies on single frequency and broadcasting [7]. Wireless sensor networks (WSN) [13] are a set of nodes that

 ISBN: 978-1-61208-203-5

can communicate with each other. Sensor nodes measure a desired physical quantity and the base station node collects data to perform processing and to connect to the wired area network. In [17], two types of sensor nodes are used: mobile sinks (to which information is sent) and sensor nodes (sensing a physical phenomenon). The intelligent vehicular systems emerged as a good candidate for benefiting from the WSN's features. WSN can be used for detecting the formation of ice over the road, and many other applications.

TABLE III. TECHNOLOGIES FOR IAVS COMMUNICATION

Characteristic	**Wifi**	**Bluetooth**	**Zigbee**
IEEE Specification	802.11x	802.15.1	802.15.4
Operating frequency	2.4 GHz ISM, 5 GHz	2.4 GHz ISM	868 MHz, 902-928MHz, 2.4 GHz ISM
Data rate	600 Mb/s	1 Mb/s	20-250 Kb/s
Nominal TX power	15-20 dBm	0-10 dBm	(-25)-0 dBm
Nominal range	30-150 m	10-100 m	10-75 m
Max # of cell nodes	2007	8	65,000
Waking up time	-	3 s	15 ms
Characteristic	WIMAX	UWB	DSRC
IEEE Specification	802.16	802.15.3	802.11.p
Operating frequency	10-66 Ghz	3.1-10.6 GHz	5.9 Ghz
Data rate	Speed up to 70Mbps	100-500 Mb/s	27 Mb/s
Nominal TX power	13.5 db	-41.3 dBm/MHz	75 MHz
Nominal range	3-5 km And 50 km from Base-stat	10 m	1000 m
Max # of cell nodes		8	
Waking up time		Narrow pulse	

The solution we propose is to use wireless sensor technology to control merchandise in this phase. The technologies integrated in the nodes include some wireless sensor device (GPS, sensors and clock) which make it possible to control the operations in real time. This allows sending relevant alerts in some critic situations and report emergency situations. UWB spectrum 3.1-10.6 GHz supports high data rate communications up to 480Mbps at a short distance (10-15m). Table III summarizes these technology choices that can be potential candidates for IAVs based data communication system. WLAN is recommended for confined regions because of its 100m range and low mobility whereas Mobile Wimax (802.16.x) is good candidate for high rate of data transfer it supports high speed transmission (up to 1Mbps in high mobility and 70 Mbps in low mobility). The gain of complexity and mobility management is due to All-IP core future Networks allowing tolerance to multipath and self-interference. Currently, the IEEE MAC sublayer proposal for UltraWideBand (UWB), namely IEEE 802.15.a adopts the carrier sense multiple access/collision avoidance (CSMA/CA) technique.

V. CONCLUSION AND FUTURE WORKS

In the present work, we discussed about the role of ICT to achieve the goal of sustainable freight transport system. Various ICTs are described with their potential benefits and applications in freight transport system. A case is described on communication in port container terminal with specifying various suitable technologies. The communication among IAVs, and infrastructure is explained in a container terminal to enhance the productivity of the system. Through the container terminal application, we are convinced that WSN can be eventually incorporated into IAVs to overcome the problems associated to wired sensors.

In general, the transport system lies on the intersection of several domains, which naturally puts pressure on the transport sector to stay as sensitive to change. Although it has been a great development in this domain but still it is a long way to reach sustainability in freight transport. Implementation of this use case using a dedicated simulation language like NS2, in order to test the network communication performances is let to perspectives.

ACKNOWLEDGMENT

This work is performed in part of European Project 'Weastflows'. Weastflows is an Interreg IVB North West Europe (NWE) project funded by the European Regional Development Fund (ERDF) that aims to encourage a shift towards greener freight transport in the NWE region.

REFERENCES

[1] C-I. Liu, H. Jula, K. Vukadinovic and P. Ioannou, "Automated guided vehicle system for two container yard layouts", Transportation Research Part C 12 (Elsevier), pp. 349-368, 2004.

 ISBN: 978-1-61208-203-5

[2] S. Yong-Dong, P. Yuan-Yuan and L. Wei-Min, "The RFID application logistics and supply chain management", Research Journal of Applied Sciences, Vol. 4, No.1, pp. 57-61, 2009.
[3] W. Y. Choi, An Efficient Polling Scheme for Enhancing IEEE 802.11 PCF Protocol, FREQUENZ 59: 268–271, 2005.
[4] A. Ollo-López and M. E. Aramendía-Muneta, ICT impact on competitiveness, innovation and environment, "Telematics and Informatics" (Article in press) 2011.
[5] S. J. Barro-Torres, T. M. Fernandez-Carames, M. Gonzalez-Lopez and C. J. Escudero-Cascon, "Maritime freight container management system using RFID", The Third International EURASIP on RFID Technology, pp. 93-96, 2010.
[6] Y. Liang and X. Bai, "Design of RFID-Enabled Container Yard Management System", G. Huang et al. (Eds.): DET2009 Proceedings, AISC 66 (Springer), pp. 1751–1758, 2009.
[7] F. Qu, F. Wang and L. Yang, "Intelligent transportation spaces: vehicles, traffic, communications, and beyond", IEEE Communication Magazine, Nov. 2010. Vol. 48, No. 11, pp. 136-142, 2010.
[8] D. Mullen, "The application of RFID technology in a Port" http://www.aimglobal.org/technologies/rfid/resources/PortTech.pdf, [retrieved: July , 2011]
[9] Z. Luo, T. Zhang and C. Wang, RFID Enabled Vehicular Wireless Query for Travel Information in Intelligent Transportation System, IEEE International Conference on RFID-Technologies and Applications, 2011.
[10] K. S. Wong, I. W. Tsang, V. Cheung and J. T. Kwok Position Estimation for Wireless sensor networks; IEEE Global Telecommunications Conference, pp. 2772-2776, 2005.
[11] M. Forcolin, E. Fracasso, F. Tumanischvili and P. Lupieri, EURIDICE-IoT applied to Logistics using the Intelligent Cargo Concept, 17th International Conference on Concurrent Enterprising, Aachen, Germany, 2011.
[12] V. Boschian, M. P. Fanti, G. Iacobellis, and W. Ukovich, The Assessment of ICT Solutions in Customs Clearance Operations, IEEE International Conference on Systems Man and Cybernetics, Istanbul, 2010.
[13] C. Chong and S. Kumar, Sensor networks: Evolution, opportunities, and challenges" in Proceedings of the IEEE Vol. 91, pp. 1247-1256, 2003.
[14] EPC global Inc. The EPC global architecture framework 1.2, 2008.
[15] Y. Wang and A. Potter, "The application or real time tracking technologies in freight transport", The Third International IEEE Conference on Signal Image Technologies and Internet Based-system 2008.
[16] M. Mansouri, B. Sauser and J. Boardman "Application of System Thinking for Resilience Study in Maritime Transportation System of Systems" IEEE International Systems Conference, Vancouver, Canada, March 23- 26, 2009 .
[17] D. Taccomi, D. Miorandi, I. Carreras, F. Chiti and R. Fantacci, Using wireless sensor networks to support intelligent transport systems, Adhoc Networks Journal (8), pp. 462-473, Elsevier 2010.
[18] M. I. Skolnik, Radar Handbook, New York: McGraw-Hill, 2007.
[19] F. Reclus and K. Drouad "Geofencing for fleet and freight management" 9th International Conference on Intelligent Transport Systems, Telecommunications, Lille 2009.
[20] L. Chen, T. Alfred and C. Wang, "A highway freight transport plate-form for the Chinese freight Market" IEEE Forum on Integrated and Sustainable Transport Systems, Vienna, Austria, June 29, 2011.
[21] A. M. Zanni and A. L. Bristow, Emissions of CO2 from road freight transport in London: Trends and policies for long run reductions, ENERGY Policy 38, pp. 1774-1786, Science Direct 2010.
[22] G. Zacharewicz, J. C. Deschamps and J. Francois, Distributed simulation platform to design advanced RFID based freight transportation systems. Computers in Industry 62, pp. 597–612, 2011.
[23] A. E. Coronado, C. S. Lalwani, E. S. Coronado and S. Cherkaoui, Wireless Vehicular Networks to Support Road Haulage and Port Operations in a Multimodal Logistics Environment, IEEE International Conference on Service Operations and Logistics and Informatics, Beijing, 2008.
[24] http://www.weastflows.eu [retrieved: June, 2012]
[25] www.libelium.com [retrieved: June, 2012]

 ISBN: 978-1-61208-203-5

Performance Analysis of Synchronization for an OFDMA System

Jihyung Kim, Jung-Hyun Kim, Kwang Jae Lim, and Dong Seung Kwon
Mobile Convergence Research Division
ETRI
DAEJEON, KOREA
Email: {savant21, jh.kim06, kjlim, dskwon}@etri.re.kr

***Abstract*—We present the time and frequency synchronization algorithm as well as the cell search scheme for cellular systems. Coarse time and fractional frequency offset estimates are performed in the time domain by using the primary preamble, while the cell search and the integer frequency offset estimate are performed in the frequency domain and afterwards the fine time and frequency offset estimates are performed by using the secondary preamble. All algorithms are evaluated under rapidly time-varying multipath fading channels and an initial carrier frequency mismatch. The simulation results show that the proposed algorithm can provide a robust synchronization and cell-search capability, even in bad cellular environments.**

Keywords-synchronizatoin; preamble; OFDMA.

I. INTRODUCTION

Orthogonal frequency division multiplexing access (OFDMA) is an efficient multicarrier technique which has been proposed for current and next generation wireless communication systems, for example IEEE 802.16e (mobile WiMAX) and IEEE 802. 16m [1] [2]. However, time offsets (TOs) and frequency offsets (FOs) between the base-station (BS) and the mobile users (MUs) arising due to local oscillator mismatch and/or the mobility of the users destroy the orthogonality among the user's sub-carriers. Inter symbol interference (ISI) and Inter carrier interference (ICI) caused by these TOs and FOs severely degrade the performance of the whole system. Therefore, the estimation and the compensation of TO and FO are imperative at the BS and the MUs.

In many situations that time delays and Doppler spreads exist, TO and FO synchronization only at the downlink of OFDMA systems is not sufficient and calls for TO and FO synchronization at the uplink as well. However, TO and FO synchronization is much more challenging at the uplink due to the presence of multiple TOs and FOs and the fact that the received signal at the BS is the sum of the transmitted signals from all the users. In addition, ensuring the identifiability of cell is also an important requirement. By utilizing cell-specific reference signal, target TO and FO can be restored from a signal disturbed by other users' TOs and FOs mismatch.

Various solutions of synchronization for OFDMA systems have been proposed in the literature [3]$\sim$ [6], but only a few of them also discuss the identifiability [7] [8]. In this paper, we invest an overall synchronization process including the identification. Moreover, we consider the mobility. The mobility of the receiver relative to transmitter is the main factor that affects the rate of fading. As the receiver moves with some velocity relative to the transmitter, the phase shifts of the received signal changes. This phenomenon is known as the Doppler effect. In practical OFDMA systems, a frequency offset due to the Doppler effect usually exists between the transmitter and the receiver.

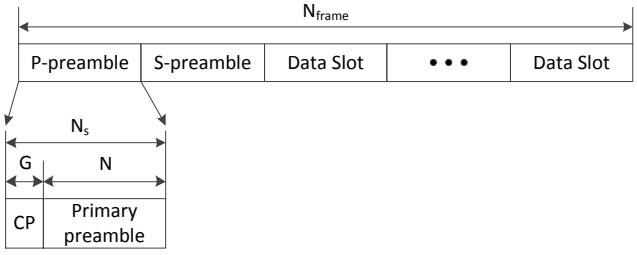


Figure 1. Abstract frame structure

This paper is organized as follows: Section II introduces a basic OFDMA system. Section III presents the proposed synchronization algorithm in detail. Section IV shows the performance of overall proposed scheme and comparison with the performance of without-fine synchronization scheme. Finally, Section V gives some conclusions.

II. SYSTEM MODEL

As shown in Figure 1, we consider a packet-based OFDMA communication system, where a preamble is placed at the beginning of the packet. The preamble consists of two different components, which are denoted as the primary preamble and the secondary preamble as similar to IEEE802.16m [2]. The length of each preamble is an OFDM symbol, which is the same length of a data symbol. We consider an OFDM symbol with two identical halves in the time domain as the primary preamble. The primary preamble is common for all BSs and is used for the TO and FO synchronization. The secondary preamble is used for the cell-ID identification and it can also be used to estimate more accurate TO and FO in bad cellular environments. In this paper, we propose a robust synchronization technique

 ISBN: 978-1-61208-203-5

using the primary preamble for coarse synchronization and using the secondary preamble for the fine synchronization as well as the cell-ID identification.

In the time domain, the n-th sample of a base-band equivalent OFDM symbol is given by

$$x(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X(k) e^{j2\pi kn/N}, \quad (1)$$

where $j = \sqrt{-1}$, $-G \leq n \leq N-1$, N is the total number of subcarriers for an OFDM symbol, $X(k)$ is the k-th modulated signal in the frequency domain, and G is the length of cyclic prefix (CP), which is assumed to be longer than the length of channel impulse response. The signal is transmitted through a frequency selective channel. Let $h(n)$ denote the base-band equivalent discrete-time channel impulse response of length ν. A carrier frequency offset of ϵ (normalized with subcarrier spacing) causes a phase rotation of $2\pi\epsilon n/N$. Assuming a perfect sampling clock, the received samples of the OFDM symbol are given by

$$y(n) = e^{j[(2\pi\epsilon n/N)+\epsilon_0]} \sum_{l=0}^{\nu-1} h(l)x(n-l) + z(n), \quad (2)$$

where ϵ_0 is an initial arbitrary carrier phase and $z(n)$ is a zero mean symmetric complex white Gaussian noise with variance σ_z^2. $x(n-l)$ is the $(n-l)$-th transmitted sample in the time domain.

III. Robust Synchronization and Cell Search Scheme

We construct a procedure including coarse time and frequency synchronization, cell search, and fine time and frequency synchronization as illustrated in Figure 2. If the decoder fails several times by a wrong cell-ID or large synchronization errors, the synchronization will be refreshed. Similarly when the estimated cell is not confirmed after several iterations, which means that the cell were estimated falsely, the complete process is performed from the beginning. We now introduce the procedure of overall synchronization scheme step by step.

A. Coarse Time and Frequency Offset Estimation

For coarse synchronization, the primary preamble is used. The primary preamble excluding CP consists of two identical halves. Using the character of the primary preamble, we can obtain a coarse TO and FO jointly. Based on [9], the timing metric by using auto-correlation in the time domain is given by

$$M(k) = \frac{|P(k)|^2}{R^2(k)}, \quad (3)$$

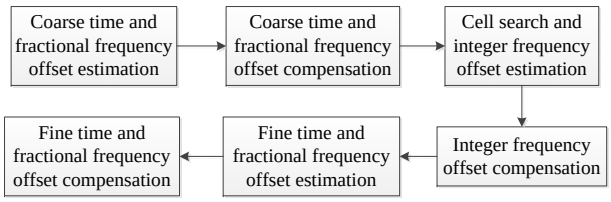


Figure 2. Proposed synchronization process

where $P(k)$ and $R(k)$ are

$$P(k) = \sum_{i=0}^{G+N/2-1} y(k+i)y^*(k+i+N/2), \quad (4)$$
$$R(k) = \frac{1}{2} \sum_{i=0}^{G+N-1} |y(k+i)|^2,$$

where $(\cdot)^*$ denotes the complex conjugate, $y(k)$ is the k-th sample of the base-band equivalent received signal, and $R(k)$ gives an estimate of the energy in $G+N$ samples of the received signal for the normalization. $N/2$ is chosen such that the angle of $P(k)$ lies in the range [-π,π] [10].

Then, we can obtain the estimated TO, τ_{coar}, and the estimated FO, ε_{coar}, from the metric $M(k)$ and $P(k)$, separately.

$$\tau_{coar} = arg \max_k M(k), \quad (5)$$
$$\varepsilon_{coar} = -\frac{1}{\pi} \angle(P(k)).$$

B. Cell search and integer FO estimation

The cell-ID can be jointly estimated with the integer FO. For the cell search, a simple method is to use a cross correlation between the received signal and the secondary preamble in the frequency domain. However, the algorithm based on the cross correlation between the reference signal and the received signal of the secondary preamble has performance degradation due to TO and FO mismatch between them. Therefore, we adopt a differential cross-correlation method that is robust to TO and FO over frequency selective channels and has the clear peak value at the estimated frame timing from the coarse synchronization.

For the integer FO estimation, we exploit the cyclic shifts of the secondary preamble sequence according to the cell-ID. The quantity of integer FO can be estimated by the subcarrier cyclic shift value maximizing the metric. This is described as follows:

$$C(k,i) = \frac{W(k,i)}{V(k,i)}, \quad (6)$$

 ISBN: 978-1-61208-203-5

where

$$W(k,i) = \sum_{k=0}^{N_l-D-1} \alpha_{k+D}\alpha_k^*\beta_{k+D,i}^*\beta_{k,i}, \quad (7)$$

$$V(k,i) = \frac{1}{4\eta}\sum_{k=0}^{N_l-D-1} |\alpha_{k+D}|^2 + |\alpha_k|^2 + \eta|\beta_{k+D,i}|^2 + \eta|\beta_{k,i}|^2, \quad (8)$$

$\alpha_k = Y(L_c(k))$, α_{k+D}=$Y(L_c\ (k+D))$, $\beta_{k,i}$=$S((L_c(k)+i)modN)$, and $\beta_{k+D,i}$=$S((L_c(k+D)+i)modN)$. η is $\frac{1}{N_l}\sum_{k=0}^{N_l-D-1}Y(L_c(k))$ which is the normalization factor. $Y(\cdot)$ is the received signal and $S(\cdot)$ is the reference signal, and $L_c(\cdot)$ is a pilot subcarrier index allocated in the frequency domain. D is the coefficient for the differential cross-correlation and set to 1 in this paper.

Then, we can obtain the estimated cell-ID, κ, and the estimated integer FO, ξ, from the metric $C(k,i)$.

$$\kappa = arg\max_k C(k,i), \quad (9)$$
$$\xi = arg\max_i C(k,i).$$

The magnitude $|C(k,i)|$ is expected to show a centrally located high peak over threshold when the cell-ID was correctly estimated. In the opposite case, if a correlation will be performed between two unequal reference sequences or between a reference sequence and received signal, then $|C(k,i)|$ will not show a significant peak and the estimated cell-ID shall be considered as false.

C. Fine Time and Frequency Offset Estimation

After performing the coarse time synchronization scheme, the estimated time is moved earlier by a few samples. This means that the FFT starting point is located within CP, and then makes to avoid ISI. The conventional method applied to fine time synchronization calculates a cross correlation value generated by producing the received signal and local samples with known long training symbols. However, the multipath channel introduces inter-path interference (IPI) into the received signal which can not be removed by the correlation based method in the conventional time offset estimation. Thus, we propose a new fine time and frequency synchronization scheme by using the secondary preamble.

Assuming a candidate of FO as ϵ in the fixed point, we can calculate $j(n)$ from the received signal $y(n)$ as follows:

$$j(n) = e^{-2\pi\epsilon n}y(n). \quad (10)$$

For a sake of simplicity, $\mathbf{j}$ is a vector of $j(\cdot)$ in the time domain and $\mathbf{S}$ is a vector of the reference signal $S(\cdot)$ in the frequency domain.

$$\mathbf{J} = FFT(\mathbf{j}). \quad (11)$$

Then,

$$\begin{aligned}\mathbf{h} &= IFFT(\mathbf{J}\cdot\mathbf{S}^*), \quad (12)\\ &= \frac{1}{N}\sum_{k=0}^{N-1} Y(k)S^*(k)e^{j2\pi kn/N}\\ &= \frac{1}{N}\sum_{k=0}^{N-1} X(k)H(k)e^{j2\pi k\delta/N}S^*(k)e^{j2\pi kn/N}\\ &= \frac{1}{N}\sum_{k=0}^{N-1} H(k)e^{j2\pi k(n+\delta)/N}\\ &= h(n+\delta)\end{aligned}$$

where $X(k)S^*(k)$=1 in the case of the corrected cell-ID estimation and the $\mathbf{h}$ is the channel impulse response vector. Then, we can jointly obtain the estimated TO, τ_{fine}, and the estimated FO, ε_{fine}, from the final metric, $h^{'}(\cdot)$.

$$\tau_{fine} = \max_k h^{'}(k,\epsilon), \quad (13)$$
$$\varepsilon_{fine} = arg\max_\epsilon h^{'}(k,\epsilon),$$

where

$$h^{'}(k,\epsilon) = \sum_{k=0}^{H-1} h(k,\epsilon). \quad (14)$$

Here, $h^{'}(\cdot)$ is calculated by a summation of the first metric $h(\cdot)$ with length H in order to improve the performance of the peak value in the multi-path fading channel.

IV. Simulation Results

Several simulations are carried out to evaluate the performance of the proposed synchronization method. The main parameters for an OFDMA system are chosen as follows: The number of subcarriers is 2048 and the nominal channel bandwidth is 20MHz, and the carrier frequency is 2.4GHz. The used channel is Veh-A of ITU-R of which the relative delay and the average power are [0;310;710;1090;1730;2510](ns) and [0;-1;-9;-10;-15;-20](dB), respectively [11].

The performance achieved by the proposed estimator is evaluated in terms of the mean squared error(MSE) according to signal-to-noise ratio(SNR), $E[|H(k)S(k)|^2/|Z(k)|^2]$. It is computed as

$$MSE(\hat{\theta}) = E[|\hat{\theta}-\theta|^2], \quad (15)$$

where $E[\cdot]$ denotes the expectation and $\hat{\theta}$ is an estimated value with respect to θ.

Figures 3 and 4 depict the MSE curve when the time offset by propagation delay is $500\mu s$ and the normalized frequency offset is 1.01. We can see that the proposed algorithm works robustly for a high delay spread fading channel in the presence of both Doppler spread and an initial frequency offset.

 ISBN: 978-1-61208-203-5

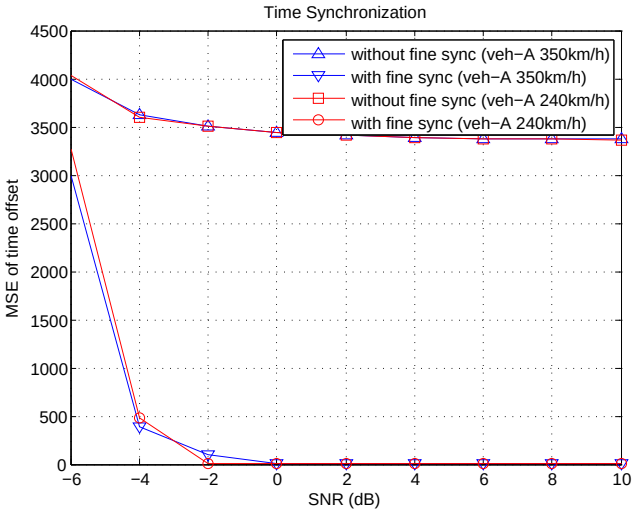


Figure 3. Mean Squared Error of time synchronization

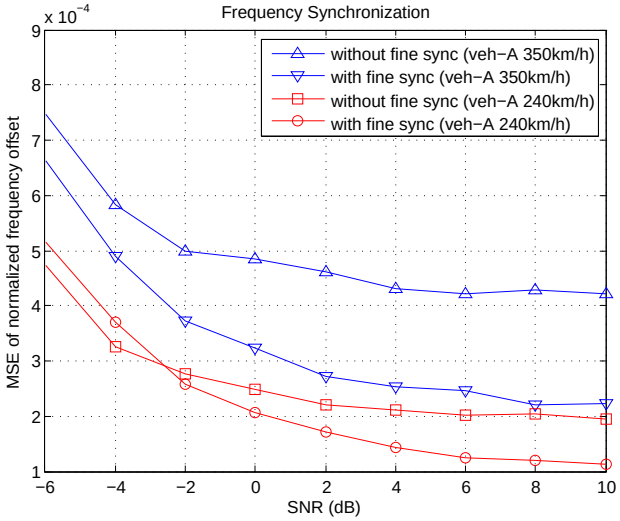


Figure 4. Mean Squared Error of frequency synchronization

Figure 5 shows the code detection probability for cell search. It provides the reliable cell search ability without additional reference signal by the result of time synchronization. Furthermore, the simulation results show that the performance of the proposed estimator is tolerant to the variation in the mobility of a user.

V. Conclusion

In this paper, a novel preamble-based synchronization and cell-search technique for OFDMA cellular systems was proposed. The preamble is composed of the primary and the secondary preamble. With the primary preamble, the initial coarse time and frequency offset estimation are performed. With the secondary preamble, the cell-search algorithm is proceeded in a hierarchical manner, the integer frequency offset estimation and the cell-ID sequence estimation. The fine time and frequency offset estimation is also performed with the secondary preamble. The overall performance of the synchronization and cell search was analyzed in terms of MSE in time and frequency-selective fading channels. Therefore, we conclude that the proposed algorithm provides the robust synchronization and cell-search capability, even in bad cellular environments.

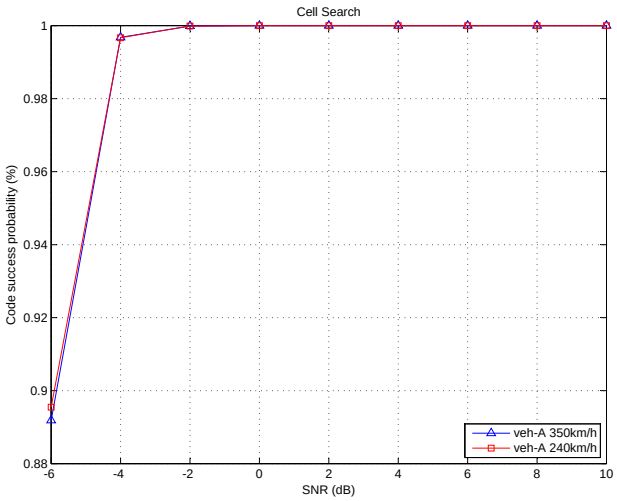


Figure 5. Success probability of cell search

Acknowledgment

This research was supported by the KCC (Korea Communications Commission), Korea, under the R&D program supervised by the KCA (Korea Communications Agency) (KCA 10-911-04-003)

References

[1] IEEE Std 802.16™-2009, "Part 16: Air interface for broadband wireless access systems," *IEEE*, May 2009.

[2] IEEE Std 802.16m™ 2011, "Part 16: Air interface for broadband wireless access systems," *IEEE*, May 2011.

[3] T. Pollet, M. Van Bladel, and M. Moeneclaey, "Ber sensitivity of ofdm systems to carrier frequency offset and weiner phase noise," *IEEE Transactions on Communications*, vol. 43, pp. 191–193, February 1995.

[4] Po-Sen Wang, Kai-Wei Lu, D.W.Lin, and Pangan Ting, "Quasi-maximum likelihood initial downlink synchronization for IEEE 802.16m," *SPAWC2011*, pp.521-525, June 2011.

[5] T.M.Schmidl and D.C.Cox, "Robust frequency and timing synchronization for ofdm," *IEEE Transactions on Communications*, vol. 45, pp. 1613–1621, December 1997.

[6] Y.Mostofi and D.C.Cox, "Robust timing synchronization design in ofdm systems - part ii: high-mobility cases," *IEEE Transactions on Wireless Communications*, vol. 6, pp. 4340–4348, December 2007.

 ISBN: 978-1-61208-203-5

[7] Dong-Uk Lee, Pansoo Kim, and Wonjin Sung, "Robust frame synchronization for low signal-to-noise ratio channels using energy-corrected differential correlation," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, 2009.

[8] K.Taura, M.Tsujishita, M.Takeda, H.Kato, M.Ishida, and Y.Isida, "A digital audio broadcasting (dab) receiver," *IEEE Transactions on Consumer Electronics*, vol. 42, pp. 322–327, Augest 1996.

[9] Chi-Min Li, Wei-Tse Sun, and Pao-Jen Wang, "An overlap s&c method for ofdm synchronization," *IEICE Electronics Express*, vol. 7, pp. 1773–1777, November 2010.

[10] K.Vasudevan, "Digital Communications and Signal Processing," Universities Press 2010, 2nd ed., www. universitiespress.com, 1996.

[11] ITU-R M.1225, "Guidelines for the evaluation of radio transmission technologies for imt-2000," *ITU-R*, 1997.

 ISBN: 978-1-61208-203-5

Mobility Load Balancing Scheme based on Cell Reselection

Toshiaki Yamamoto, Toshihiko Komine, and Satoshi Konishi
KDDI R&D Laboratories Inc.
2-1-15 Ohara, Fujimino-shi, Saitama, 356-8502 Japan
{tos-yama, to-komine, skonishi} @kddilabs.jp

***Abstract*—Mobility Load Balancing (MLB) is one of the most important functions of Self-Organizing Networks (SON) in Long Term Evolution (LTE). The conventional MLB schemes based on handover (HO) conflict with Mobility Robustness Optimization (MRO) because both operations adjust the same HO parameters. The simulation results show that the conventional MLB scheme cannot achieve load balancing gain without some degradation in HO performance. In order to solve the conflict problem, this paper proposes an MLB scheme based on cell reselection (CR) that works in coordination with MRO. The proposed scheme adjusts the CR parameters and not the HO parameters, and never conflicts with MRO. Through computer simulations, it is shown that the proposed scheme can realize effective load balancing on a par with conventional schemes without any degradation in HO performance. The simulation results show that the proposed scheme is especially effective in an environment where a lot of small-size data packets are transmitted by a large number of users, which is highly applicable to current mobile networks with explosive diffusion of smart phones. In such case, more than 10% and 90% gains can be obtained in the total throughput and 5th percentile user throughput, respectively.**

Keywords: LTE; self-organizing networks (SON); mobility load balancing (MLB); mobility robustness optimization (MRO); cell reselection

I. INTRODUCTION

In conventional cellular networks, system parameters are manually adjusted to maintain and/or improve the operational performance. However, due to the rapid evolution of networks, the parameters have become more complex and larger, and such manual tuning of the parameters is becoming increasingly difficult. In order to reduce the operational complexity of cellular networks, the concept of Self-Organizing Networks (SON) has been introduced into Long Term Evolution (LTE) and is currently being discussed in the 3rd Generation Partnership Project (3GPP) [1].

One of the main functions of SON is Mobility Load Balancing (MLB) [2]. In cellular networks, traffic demand dynamically changes both in time and space, and it is common for some cells to be heavily loaded, whereas their adjacent cells are not. The objective of MLB is to distribute cell load evenly among adjacent cells or to transfer part of the traffic from congested cells. In MLB, this is done by self-optimization of the mobility parameters.

In the 3GPP, the concept, requirements, procedures and interfaces of MLB are discussed. The actual solutions are left to vendor specific algorithms and several algorithms for the optimization of the mobility parameters have been reported in the literature [3][4][5][6][7][8][9][10][11]. They considered MLB based on handover (HO), which is referred to as "HO-MLB" hereafter. HO-MLB adjusts the HO timing by biasing the HO measurements, forcing user equipments (UEs) around the cell-edge in highly loaded cells to hand off to less loaded neighboring cells in order to share traffic between adjacent cells. The unavoidable problem in any attempt to realize HO-MLB is the conflict with Mobility Robustness Optimization (MRO), which is also one of the important functions of SON [2]. MRO aims to minimize HO failures and reduce ping-pong HOs by adjusting the HO parameters. Therefore, it is possible that both HO-MLB and MRO adjust the same HO parameter in the opposite directions at the same time and the conflict may lead to performance degradation. The conflict problem between HO-MLB and MRO was investigated in detail [6]. In [6], the authors proposed a solution to avoid the conflict problem by imposing a restriction on HO-MLB through setting an allowed range to make sure that HO failures never occur. However, the proposal is not a sufficient solution because the load balancing is extremely limited by the operation of MRO and the gain is therefore expected to be negligible. Moreover, it is very difficult to determine the allowed range for in-service cellular networks where the allowed range varies dynamically in response to changes in the radio environment and UE mobility. In such networks, the scheme may not work effectively.

Cell reselection (CR) also has the potential to realize load balancing [2]. We refer to an MLB based on CR as "CR-MLB" hereafter. In the same way as HO-MLB, the adjustment of the CR timing by biasing the CR measurements causes the UEs around the cell-edge in highly loaded cells to migrate to less loaded neighboring cells. While HO-MLB is intended for UEs in radio resource control (RRC) connected mode, CR-MLB is intended for UEs in RRC idle mode. In previous studies of MLB, HO-MLB is mainly studied because the adjustment of CR parameters is only effective during call set-up and the optimization of HO parameters is considered to be the preferred option [9]. However, CR-MLB has a great advantage over HO-MLB in that there is no conflict with MRO because CR-MLB and MRO adjust different parameters. Therefore, CR-MLB can be a promising way to realize load balancing more than HO-MLB. To the best of our knowledge, the performances of CR-MLB have not been reported.

 ISBN: 978-1-61208-203-5

In this paper, we propose a CR-MLB scheme that works in coordination with MRO. The proposed scheme adjusts CR parameters under the restriction determined by the HO parameters, which are obtained by MRO operating independently of MLB. The proposed scheme is expected to realize load balancing without any degradation in HO performance. In order to examine the performance and clarify the applicable scope of the proposed scheme, we perform the computer simulations over different UE distributions and traffic patterns.

The rest of the paper is organized as follows. In section II, we explain the conventional HO-MLB scheme, and the influence on HO performance is clarified through a computer simulation. Section III introduces the proposed CR-MLB scheme. The performance of the proposed scheme is evaluated in section IV. Finally, we conclude this paper in section V.

II. MOBILITY LOAD BALANCING BASED ON HANDOVER

In this section, we introduce the HO procedure in 3GPP LTE and then explain the operational principle of the conventional MLB scheme based on the HO (HO-MLB). The influence of HO-MLB on HO performances is also examined by computer simulations.

A. HO Procedure

The HO procedure in 3GPP LTE begins with measurement report (MR) transmission from a UE to its source cell (serving cell). The UE periodically performs downlink channel measurements based on cell-specific reference signals and checks whether the signal strengths satisfy the conditions for MR transmission. The entering condition for event A3, which is one of the MR triggering events generally used for the HO procedure [12], is defined as

$$M_1 - M_0 > Hys_0 + a3Offset_0 - CIO_{0,1}, \quad (1)$$

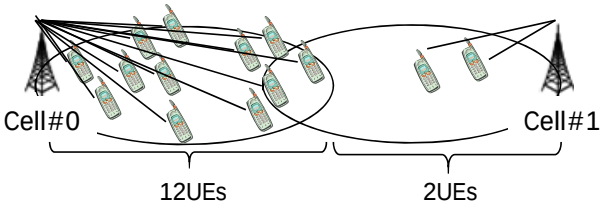
where M_0 and M_1 are the signal strengths of the serving cell (Cell#0) and the target cell (Cell#1), respectively. Hys_0 is the hysteresis parameter and $a3Offset_0$ is the offset parameter for event A3. In order to simplify the discussion without generality, we disregard Hys_0 and $a3Offset_0$ in this paper. By substituting $Hys_0 = a3Offset_0 = 0$ into (1), we obtain the following condition;

$$M_1 - M_0 > -CIO_{0,1}, \quad (2)$$

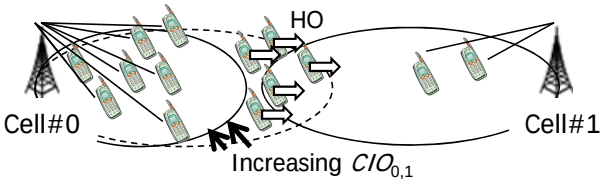
where $CIO_{0,1}$ is a cell-specific offset parameter set by Cell#0 for Cell#1 and is called "Cell Individual Offset (CIO)." If condition (2) is satisfied for duration of Time To Trigger (TTT), the UE sends the MR to Cell#0 and the HO procedure from Cell#0 to Cell#1 is initiated.

B. Conventional HO-MLB Scheme

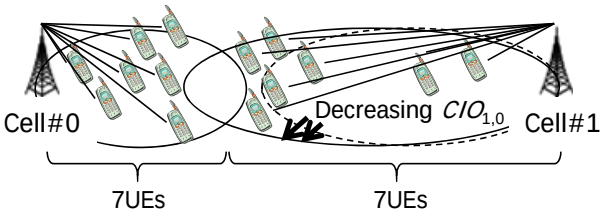
The operational principle of HO-MLB is illustrated in Figure 1. As shown in Figure 1(a), 12 UEs and 2 UEs belong to Cell#0 and Cell#1, respectively, that is, Cell#0 is more loaded than Cell#1. In this case, as illustrated in Figure 1(b), the HO-MLB scheme increases $CIO_{0,1}$ to make the HO timing from Cell#0 to Cell#1 earlier and UEs of Cell#0 near Cell#1 will hand off to Cell#1. In addition, as shown in Figure 1(c), the HO-MLB scheme can decrease $CIO_{1,0}$ to make the HO timing from Cell#1 to Cell#0 later in order to keep the handed off UEs from Cell#0 staying in Cell#1. Finally, the load is shared equally between Cell#0 and Cell#1.



(a) Before the operation of HO-MLB.



(b) HO-MLB Increases $CIO_{0,1}$ to force UEs around cell-edge to hand off to Cell#1.



(c) HO-MLB decreases $CIO_{1,0}$ to keep UEs around the cell-edge staying in Cell#1.

Figure 1. Operational principle of HO-MLB.

C. Influence on HO pereformances

The adjustment of the HO timing by the HO-MLB scheme may cause performance degradation of HO processes. Here the influence of HO-MLB on HO performances is examined through computer simulations. We evaluate the HO failure rate and ping-pong HO rate between Cell#0 and Cell#1 by changing $CIO_{0,1}$ and $CIO_{1,0}$. In the simulations, the number of HO failures is counted if "Too Late HO," "Too Early HO," or "HO to Wrong Cell" is observed [2]. The number of ping-pong HOs is counted when the UE returns to the original serving cell within a pre-determined minimum-time-of-stay (MTS) after a HO from the original serving cell to a neighboring cell [13] and the MTS value in the simulations is set to 2 seconds. The HO failure rate and ping-pong HO rate is defined as the ratio of the number of HO failures and ping-pong HOs divided by the number of all HOs including HO failures, respectively. The other simulation conditions are summarized in Table I.

 ISBN: 978-1-61208-203-5

In order to obtain significant load balancing gain, the MLB scheme should be operated in an environment in which low mobility UEs are dominant. Therefore, we assume that UE mobility is 3 km/h throughout the paper. In the case of 3 km/h, we run the MRO [14] and find that the optimal CIO value that minimizes the sum of the HO failure rate and ping-pong HO rate is –6 dB. The default CIO value is set to –6 dB.

TABLE I. SIMULATION CONDITIONS

Inter-Site Distance	500 m
eNB Power	43 dBm
Pathloss	120.9+37.6 $\log_{10}$(d)
Shadowing	Standard deviation: 8 dB, Correlation distance: 50 m
Fading	Typical Urban 6path
UE Mobility	Uniform Distribution, Random Walk, 3/30/60/120/240 km/h
L3 Filter Parameter	*K*: 4
Handover	T300: 1 s, T301: 500 ms, T304: 400 ms, T311: 10 s, delay (X2): 60 ms, delay (intra-eNB): 10 ms, Tstore_ue_context: 1 s, Time to Trigger: 256 ms
RLF Detection	Qin: –6 dB, Qout: –8 dB, N310: 1, N311: 1, T310: 1 s

Figure 2 shows the HO failure rate and ping-pong HO rate when the HO-MLB scheme increases $CIO_{0,1}$ from the default value of –6 dB to 6 dB while $CIO_{1,0}$ is fixed to –6 dB. It is found that the HO failure rate is almost zero for all CIO values, but the ping-pong HO rate becomes higher as the HO-MLB scheme increases $CIO_{0,1}$. This is because the hysteresis region becomes narrower by increasing $CIO_{0,1}$. It is undesirable for the incidence of ping-pong HO to be too high as it consumes a lot of radio resources as well as placing an unnecessary burden on the hardware units of eNBs and wastes backhaul resources. If we introduce a policy whereby the ping-pong HO rate is kept below 15 %, HO-MLB can increase $CIO_{0,1}$ only up to –2 dB and this may result in little load balancing gain being achieved. The gain will be evaluated in detail in section IV.

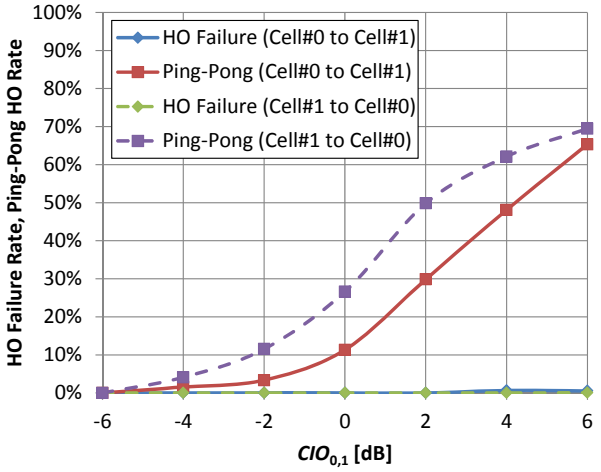


Figure 2. HO failure rate and ping-pong HO rate: $CIO_{0,1}$ is changed from –6 dB to 6 dB, while $CIO_{1,0}$ is the fixed value of –6 dB.

In order to prevent ping-pong HO from occurring, HO-MLB has to keep the hysteresis region at 12 dB by decreasing $CIO_{1,0}$ according to the increase in $CIO_{0,1}$ as shown in Figure 1(c). Figure 3 shows that decreasing $CIO_{1,0}$ causes a large number of Too Late HO from Cell#1 to Cell#0 and the HO failure rate becomes critically high as HO-MLB increases $CIO_{0,1}$. It is found that the increase in HO failures is inevitable when HO-MLB decreases $CIO_{1,0}$. As a result, this option is not an appropriate approach because the HO failures directly influence the user performance.

Finally, from Figures 2 and 3, it can be concluded that HO-MLB cannot achieve load balancing gain without some degradation in HO performance.

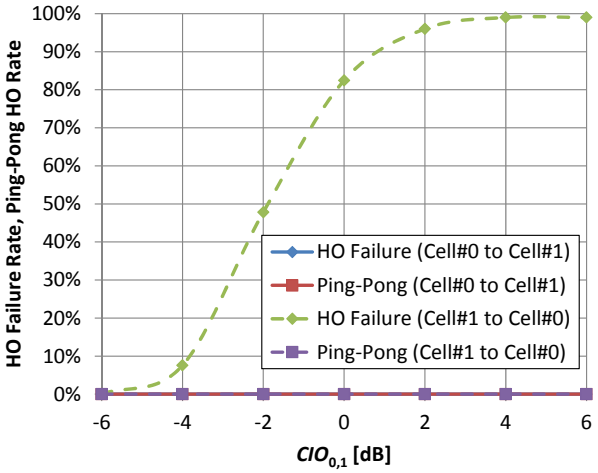


Figure 3. HO failure rate and ping-pong HO rate: $CIO_{0,1}$ is changed from –6 dB to 6 dB and $CIO_{1,0}$ is changed so that the hysteresis region is kept at 12 dB. For example, $CIO_{1,0}$ is set to –6 dB and –18 dB when $CIO_{0,1}$ is set to –6 dB and 6 dB, respectively.

III. MOBILITY LOAD BALANCING BASED ON CELL RESELECTION

In this section, we introduce the CR procedure in 3GPP LTE and then explain the operational principle of the proposed MLB scheme based on the CR (CR-MLB).

A. Cell Reselection Procedure

The CR procedure in 3GPP LTE is performed as follows. The UE in RRC idle mode periodically performs idle mode measurements. When more than 1 second has elapsed since the UE camped on the current serving cell and the following condition (3) is satisfied during a time interval of $Treselection_{RAT}$, the UE camped on Cell#0 shall perform cell reselection to Cell#1 [15].

$$M_1 - M_0 > Q_{Hyst,0} + Qoffset_{0,1}, \quad (3)$$

where M_0 and M_1 are the signal strengths of the camped cell (Cell#0) and the target cell (Cell#1), respectively. $Q_{Hyst,0}$ is the hysteresis parameter and, to simplify the discussion, we

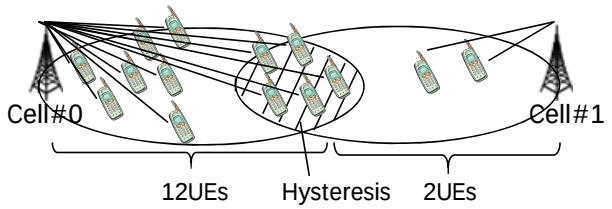
disregard it in this paper. By substituting $Q_{Hyst,0} = 0$ into (3), we obtain the following condition;

$$M_1 - M_0 > Qoffset_{0,1}, \quad (4)$$

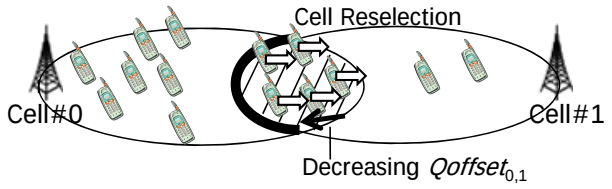
where $Qoffset_{0,1}$ is a cell-specific offset parameter set by Cell#0 for Cell#1 called Qoffset. In normal operation, the CR timing is adjusted to be the same as the HO timing and then $Qoffset_{0,1}$ is equal to $-CIO_{0,1}$.

B. Proposed CR-MLB Scheme

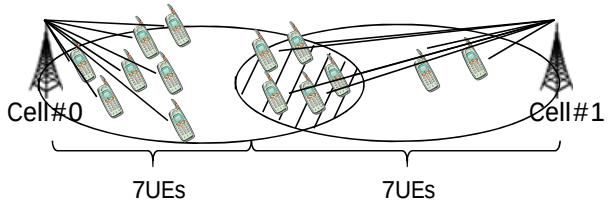
Figure 4 illustrates the operational principle of CR-MLB. As illustrated in Figure 4(a), 12 UEs and 2 UEs in RRC connected mode belong to Cell#0 and Cell#1, respectively. Because Cell#0 is more loaded than Cell#1, the CR-MLB scheme decreases $Qoffset_{0,1}$ to make the CR timing from Cell#0 to Cell#1 earlier as shown in Figure 4(b). Once UEs of Cell#0 leave the RRC connected mode and enter the RRC idle mode, the UE near Cell#1 will perform cell reselection to Cell#1. Finally, as shown in Figure 4(c), the UEs camped on Cell#1 will enter the RRC connected mode and belong to Cell#1. In summary, if the UE stays within the hysteresis region of the HO parameter, it can potentially connect to both Cell#0 and Cell#1. Once the UE connects to either cell, the UE does not satisfy the HO condition and continues to stay in the same cell. CR-MLB moves such UEs to the less loaded cell when they are in the RRC idle mode.



(a) UEs in RRC connected mode before the operation of CR-MLB.



(b) CR-MLB decreases $Qoffset_{0,1}$ to force UEs in RRC idle mode around the cell-edge to perform cell reselection to Cell#1.



(c) UEs in RRC connected mode after the operation of CR-MLB.

Figure 4. Operational principle of CR-MLB.

In the proposed scheme, CR-MLB and MRO operate independently and never conflict with each other. MRO always operates and optimizes the setting of CIO to minimize the HO failure rate and ping-pong HO rate. When the traffic load in each cell is not heavy and it is not necessary to perform load balancing, CR-MLB does not operate and the CR parameter is set to the default value $Qoffset_{0,1} = -CIO_{0,1}$, which is determined by the result of MRO. If the traffic load in some cells becomes high and the traffic loads between adjacent cells are unbalanced, CR-MLB begins to operate and decreases the CR parameter of the higher loaded cell for the less loaded cells. As a result, CR-MLB works in coordination with MRO and can realize load balancing without any degradation in HO performance.

There is one concern that will need to be considered. When CR-MLB decreases $Qoffset_{0,1}$ and the hysteresis margin between the CR timing from Cell#0 to Cell#1 and the HO timing from Cell#1 to Cell#0 are close to zero, there is a high possibility for the UE performing cell reselection to Cell#1 in RRC idle mode to return to Cell#0 by HO after the UE becomes the RRC connected mode. This is because the measured signal strengths of Cell#0 and Cell#1 fluctuate over time due to the effect of fast fading, shadowing, and so on. Note that though this may reduce the load balancing gain of CR-MLB, it never causes ping-pong HO in RRC connected mode because the hysteresis of the HO parameters between $CIO_{0,1}$ and $CIO_{1,0}$ is not changed by CR-MLB. In the following section, we examine the load balancing performance of CR-MLB considering the above concern.

IV. Performance Evaluation

In this section, the load balancing performance of the proposed CR-MLB scheme is examined through computer simulations. The system-level simulator written in C++ is developed for the evaluation. First, we compare the proposed scheme with the conventional HO-MLB scheme for various offered traffic loads. We also evaluate the proposed scheme over different UE distributions and traffic patterns to clarify the applicable scope.

A. Simulation Conditions

Figure 5 illustrates a part of the cell layout used in the simulation. In order to examine the load balancing gain, we change CIO and Qoffset values of Cell#0 for the neighboring 6 cells (Cell#1, Cell#2, Cell#4, Cell#5, Cell#8, and Cell#19) and evaluate the downlink UE throughput in those cells.

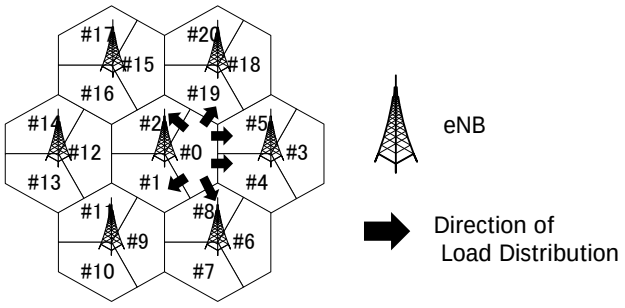


Figure 5. A part of the cell layout used in the simulation.

We define the ratio of the throughput to that for the case where $\{CIO_{0,1} = -6$ dB , $Qoffset_{0,1} = 6$ dB$\}$ as throughput gain. From the simulation results, the throughput gain for the case where $\{ CIO_{0,1} = x$ dB , $Qoffset_{0,1} = y$ dB $\}$ is derived as

$$Throughput\ gain = \frac{Throughput(CIO_{0,1} = x\text{dB}, Qoffset_{0,1} = y\text{dB})}{Throughput(CIO_{0,1} = -6\text{dB}, Qoffset_{0,1} = 6\text{dB})} - 1.0 , \quad (5)$$

and the gains of total throughput, 5th percentile user throughput, and 50th percentile user throughput are evaluated. As an example, throughput gain of 1.0 means that the throughput is twice as that for the case where $\{CIO_{0,1} = -6$ dB , $Qoffset_{0,1} = 6$ dB$\}$.

As for the traffic model, we assume that traffic with a fixed data size arrives at a fixed time interval per UE. Note that the timing of traffic arrival is uniformly distributed among UEs. When the traffic arrives, the UE enters the RRC connected mode and then returns to the RRC idle mode 1 second after the transmission traffic queue becomes empty. If the queue is not empty when the next traffic arrives, the UE continues to stay in the RRC connected mode. The other simulation conditions are summarized in Tables I and II.

TABLE II. SIMULATION CONDITIONS

Cell Layout	Hexagonal grid, 57 cells
Carrier Frequency	900 MHz
System Bandwidth	10 MHz
Scheduler	Round Robin
# of Antennas	TX: 1, RX: 2
# of UEs	20 UEs per Cell
Traffic Model	Traffic with a fixed data size at a fixed time interval Default data size: 475 KB Default time interval: 30 s
CIO	For Cell#0 to neighboring cells: variable For others: –6 dB
Qoffset	For Cell#0 to neighboring cells: variable For others: 6 dB
$Treselection_{RAT}$	0 s

B. Comparison with HO-MLB

First, we compare the load balancing performance of the proposed CR-MLB with that of the conventional HO-MLB schemes for various offered traffic loads. In order to compare them in the case where high throughput gain is expected to be obtained, we allocate 140 UEs to Cell#0 and no UEs to neighboring cells, that is, intentionally create an unbalanced traffic load situation between Cell#0 and the neighboring cells. In the CR-MLB, $Qoffset_{0,1}$ is changed from 6 to –6 dB and, in the HO-MLB, $CIO_{0,1}$ is changed from –6 to 6 dB.

Figure 6 shows the throughput gains of the CR-MLB and HO-MLB. We change the data size of traffic so that the total offered traffic load in Cell#0 is 17.6, 22, and 26.4 Mbps in Figures 6(a)-(c), respectively. The resource block (RB) usage of Cell#0 without MLB is nearly 100% when the total offered traffic load is 17.6 Mbps.

From Figure 6(a), it is found that there is little throughput gains both for the CR-MLB and HO-MLB because almost all offered traffic can be handled solely by Cell#0 even though MLB does not operate.

In Figure 6(b), we can see that high throughput gains are obtained. It is found that MLB improves the 5th percentile user throughput rather than the total throughput and about 80% gain can be achieved both with CR-MLB and HO-MLB. In HO-MLB, as $CIO_{0,1}$ increases, the throughput gains become higher although the ping-pong HO rate also becomes higher as shown in Figure 2. The throughput gains become saturated with $CIO_{0,1}$ of more than 0 dB because the overloaded offered traffic of Cell#0 is mostly transferred into neighboring cells and the RB usage of Cell#0 becomes less than 100%. In CR-MLB, as $Qoffset_{0,1}$ decreases, the throughput gains become higher without any degradation in HO performance. Compared with HO-MLB, the throughput gains of CR-MLB increase more slowly. The reason is that some UEs return to Cell#0 in the RRC connected mode due to the effect of fast fading. If the stationary UEs are dominant and the signal strengths change more slowly, it is expected that the lines of throughput gains of CR-MLB will fit closely with those of HO-MLB. In any case, CR-MLB can achieve throughput gains equivalent to HO-MLB with $Qoffset_{0,1}$ = 6 dB. In addition, as an example, if we adopt a policy where the ping-pong HO rate is kept below 15%, HO-MLB cannot increase $CIO_{0,1}$ by more than –2 dB as shown in Figure 2 and the throughput gains of CR-MLB with $Qoffset_{0,1}$ = 6 dB are superior to those of HO-MLB with $CIO_{0,1} = -2$.

Figure 6(c) shows the case where Cell#0 is very heavily loaded. In this case, the differences in the throughput gains between CR-MLB and HO-MLB become larger. However, if we adopt the policy that does not allow the ping-pong HO rate to rise above 15%, the throughput gains of CR-MLB with $Qoffset_{0,1}$ = 6 dB are comparable to those of HO-MLB with $CIO_{0,1} = -2$.

From the above observations, we can conclude that the proposed scheme can realize load balancing effectively on a par with conventional schemes without any degradation in HO performance.

C. Performance Evaluation of CR-MLB

Next, we evaluate the proposed CR-MLB scheme over different UE distributions and traffic patterns.

Figure 7 shows the throughput gains when the ratio of the number of UEs in Cell#0 to neighboring cells is changed. The number of UEs in each cell except for Cell#0 is fixed at 20 and only the number of UEs in Cell#0 is changed. The offered data size of traffic is normalized so that the total offered traffic load in Cell#0 is 26.4 Mbps. From Figure 7, we can see that the throughput gains become higher as the ratio of UEs increases and the traffic load becomes more unbalanced.

In Figure 8, we evaluate the throughput gains by changing the total offered traffic load in Cell#0. In the evaluations, we allocate 140 UEs to Cell#0 and no UEs to neighboring cells. From Figure 8, it can be seen that the throughput gains reach the maximum when the total offered traffic load is around 22 Mbps, and then they become lower as the total offered traffic load increases. This is because

 ISBN: 978-1-61208-203-5

when the total offered traffic load increases significantly and the RB usage of Cell#0 is much greater than 100%, the UEs' transmission traffic queues are always full and the UEs remain continuously in the RRC connected mode. In this case, there is no possibility of performing cell reselection and the performance gain of CR-MLB is reduced. This is a weak point of CR-MLB, but this situation can be avoided if CR-MLB operates before the RB usage of Cell#0 becomes greater than 100% and properly off-load the excess load of Cell#0 to the neighboring cells.

In Figure 9, the throughput gains are evaluated by changing the data size of traffic. The time interval of traffic arrival is adjusted according to the data size of traffic so that the total offered traffic load in Cell#0 is 26.4 Mbps. In the evaluations, we also allocate 140 UEs to Cell#0 and no UEs to neighboring cells. From Figure 9, we can see that more than 90% and 10% gains in the 5th percentile user throughput and total throughput respectively when the data size of traffic is less than 500 KB. As the data size of traffic increases, the throughput gains becomes lower because the time that a UE stays in the RRC connected mode becomes longer due to the large data size of traffic and the possibility that the UE will return to Cell#0 in the RRC connected mode becomes higher.

Finally, from Figures 7-9, we can conclude that the proposed CR-MLB scheme is especially effective in an environment where a lot of small-size data packets are transmitted by a large number of users. Note that in the simulations of Figures 7-9, we also evaluate the throughput gains of HO-MLB and confirm that those of CR-MLB are superior or comparable to those of HO-MLB for all cases if we adopt a policy where the ping-pong HO rate is kept below 15%.

V. CONCLUSION

In this paper, we have proposed a novel MLB scheme based on CR that works in coordination with MRO. The conventional MLB schemes based on HO conflict with MRO and the simulation results show that the conventional MLB scheme cannot achieve load balancing gain without some degradation in HO performance. The proposed scheme adjusts the CR parameters but not the HO parameters, and never conflicts with MRO. Through the computer simulations, it was demonstrated that the proposed scheme can realize load balancing effectively on a par with conventional schemes without any degradation in HO performance. The simulation results have also shown that the proposed scheme is particularly effective in an environment where a lot of small-size data packets are transmitted by a large number of users, which is highly applicable to current mobile networks with explosive diffusion of smart phones. In such case, more than 10% and 90% gains can be obtained in the total throughput and 5th percentile user throughput, respectively.

REFERENCES

[1] 3GPP standardization, "Self-organizing networks (SON) concepts and requirements (Release 9)," TS32.500 v9.0.0, Dec. 2009, http://www.3gpp.org/.

[2] 3GPP standardization, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Overall description Stage 2," TS36.300 v9.8.0, Sept. 2011, http://www.3gpp.org/.

[3] A. Lobinger, S. Stefanski, T. Jansen, and I. Balan, "Load Balancing in Downlink LTE Self-Optimizing Networks," in *Vehicular Technology Conference (VTC 2010-Spring), 2010 IEEE 71st*, May 2010, pp. 1-5.

[4] R. Kwan, R. Arnott, R. Paterson, R. Trivisonno, and M. Kubota, "On Mobility Load Balancing for LTE Systems," in *Vehicular Technology Conference (VTC 2010-Fall), 2010 IEEE 72nd*, Sept. 2010, pp. 1-5.

[5] H. Zhang, X. Qiu, L. Meng, and X. Zhang, "Design of distributed and autonomic load balancing for self-organization LTE," in *Vehicular Technology Conference (VTC 2010-Fall), 2010 IEEE 72nd*, Sept. 2010, pp. 1-5.

[6] Z. Liu, P. Hong, K. Xue, and M. Peng, "Conflict Avoidance between Mobility Robustness Optimization and Mobility Load Balancing," in *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, Dec. 2010, pp. 1-5.

[7] B. Wang, X. Wen, and W. Zheng, "A Self-Optimizing Method Based on Handover for Load Balancing," in *Information Theory and Information Security (ICITIS), 2010 IEEE International Conference on*, Dec. 2010, pp. 1026-1029.

[8] P. Muñoz, R. Barco, I. de la Bandera, M. Toril, and S. Luna-Ramírez, "Optimization of a fuzzy logic controller for handover-based load balancing," in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, May 2011, pp. 1-5.

[9] J. M. R. Avilés, S. Luna-Ramírez, M. Toril, F. Ruiz, I. de la Bandera-Cascales, and P. Muñoz-Luengo, "Analysis of Load Sharing Techniques in Enterprise LTE Femtocells," in *Wireless Advanced (WiAd), 2011*, June, 2011, pp. 195-200.

[10] A. El-Halaby and M. Awad, "A Game Theoretic Scenario for LTE Load Balancing," in *AFRICON, 2011*, Sept. 2011, pp. 1-6.

[11] L. Xu, Y. Chen, J. Schormans, L. Cuthbert, and T. Zhang, "User-Vote Assisted Self-Organizing Load Balancing for OFDMA Cellular Systems," in *Personal, Indoor and Mobile Radio Communications (PIMRC), 2011 IEEE 22nd International Symposium on*, Sept. 2011, pp. 217-221.

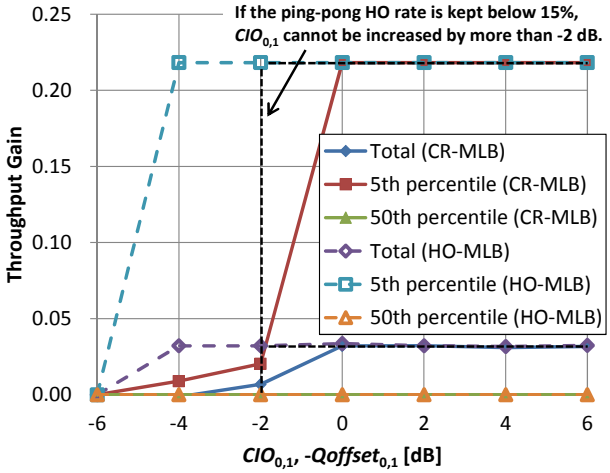
[12] 3GPP standardization "Evolved Universal Terrestrial Radio Access (E-UTRA) Radio Resource Control (RRC) Protocol specification (Release 9)," TS36.331 v9.3.0, June 2010, http://www.3gpp.org/.

[13] 3GPP standardization, "Evolved Universal Terrestrial Radio Access (E-UTRA) Mobility Enhancements in Heterogeneous Networks (Release 11)," TR36.839 v0.2.0, Sept. 2011, http://www.3gpp.org/.

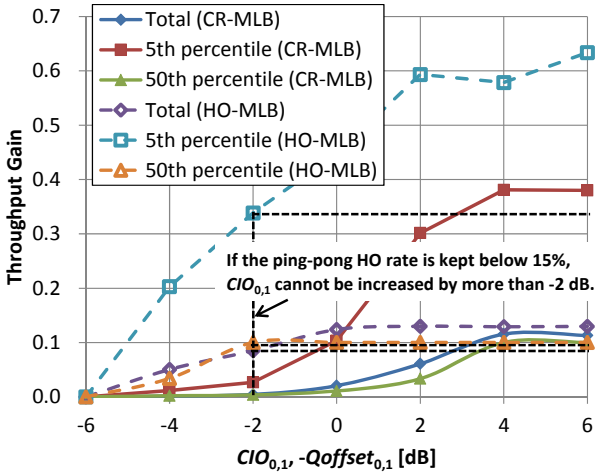
[14] K. Kitagawa, T. Komine, T. Yamamoto, and S. Konishi, "A Handover Optimization Algorithm with Mobility Robustness for LTE Systems," in *Personal, Indoor and Mobile Radio Communications (PIMRC), 2011 IEEE 22nd International Symposium on*, Sept. 2011, pp. 1647-1651.

[15] 3GPP stndardization, "Evolved Universal Terrestrial Radio Access (E-UTRA) User Equipment (UE) procedures in idle mode," TS36.304 v9.5.0, Dec. 2010, http://www.3gpp.org/.

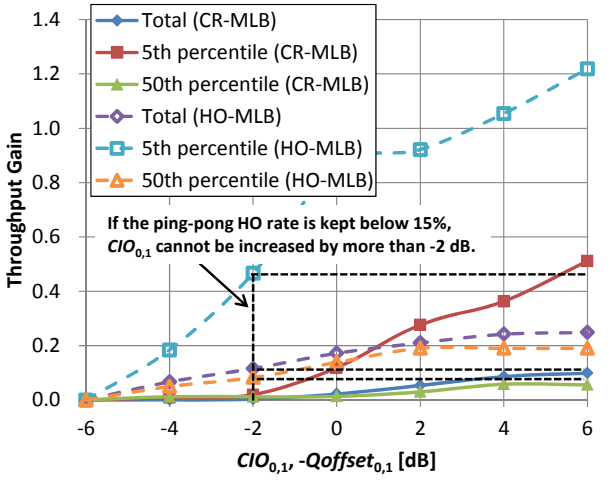
 ISBN: 978-1-61208-203-5



(a) Total offered traffic load in Cell#0 is 17.6 Mbps.



(b) Total offered traffic load in Cell#0 is 22 Mbps.



(c) Total offered traffic load in Cell#0 is 26.4 Mbps.

Figure 6. Throughput gains of the proposed CR-MLB and the conventional HO-MLB schemes: $Qoffset_{0,1}$ is changed from 6 to –6 dB in the CR-MLB and $CIO_{0,1}$ is changed from –6 to 6 dB in the HO-MLB.

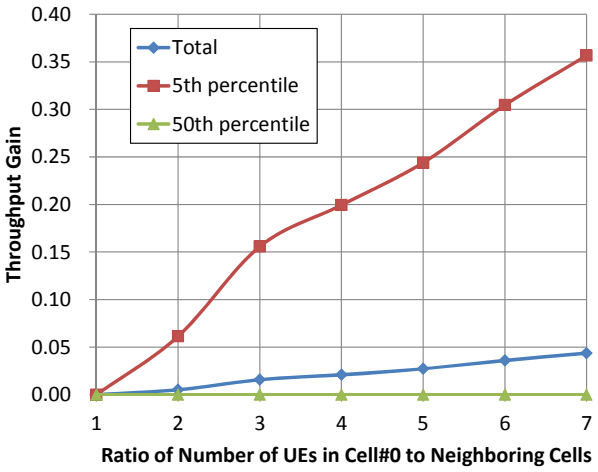


Figure 7. Throughput gains of the proposed CR-MLB scheme versus the ratio of number of UEs in Cell#0 to neighboring cells.

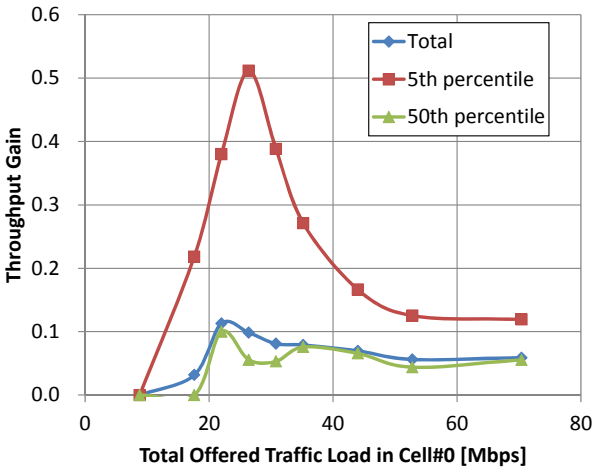


Figure 8. Throughput gains of the proposed CR-MLB scheme versus total offered traffic load in Cell#0.

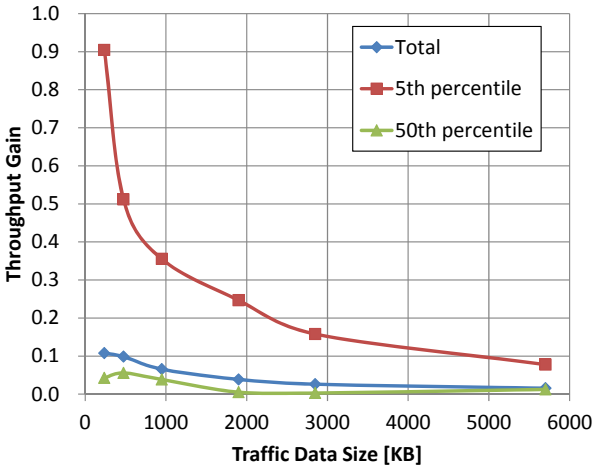


Figure 9. Throughput gains of the proposed CR-MLB scheme versus data size of traffic.

 ISBN: 978-1-61208-203-5

Trust and Energy-aware Routing Protocol for Wireless Sensor Networks

Laura Gheorghe, Răzvan Rughiniș, Nicolae Țăpuș
Department of Computer Science and Engineering
University Politehnica of Bucharest
Bucharest, Romania
{laura.gheorghe, razvan.rughinis, ntapus}@cs.pub.ro

Abstract - **Wireless Sensor Networks are composed of resource-constrained devices and are used in critical monitoring and tracking applications. Therefore, routing protocols for such networks should take into consideration the trustworthiness of and the energy available on sensor nodes. We developed TER - Trust and Energy-aware Routing protocol, a location-based, trust and energy-aware, routing protocol for Wireless Sensor Networks. The protocol uses distance, trust and energy as metrics when choosing the best path towards the destination. The protocol can be easily extended to support other metrics. We implemented our protocol in TinyOS and tested it in several test configurations. We determined experimentally that TER provides traffic load and energy balancing while building trustworthy paths.**

Keywords - Wireless Sensor Networks, routing protocol, trust, energy, security

I. INTRODUCTION

Wireless Sensor Networks (WSNs) represent an innovative technology used for monitoring specific environments. A WSN is composed of tens to thousands of sensor nodes, which are low-power, low-cost, small, resource-constrained devices. The sensor nodes collaborate in order to detect events that take place in the monitored environment and send relevant data to one or more base stations [1].

WSNs are used in critical applications like military surveillance, homeland security and medical monitoring, and, in these cases, protecting the network against malicious attacks is crucial. However, WSNs have unique characteristics: wireless transmission medium, limited resources available on sensor nodes, hostile environment, ad-hoc deployment, unreliable communication, and unattended operation. Therefore, protocols for critical sensor networks should be designed with security in mind, while taking into consideration their specific constraints and challenges.

For large sensor networks, multi-hop communication is more energy-efficient than single-hop communication. A routing protocol is used for assuring packet delivery and most network traffic has a many-to-one pattern because all nodes send data packets towards the base station.

Most routing protocols for sensor networks use a single metric to determine the best path to destination. Some use two metrics such as location and energy [2], [3], location and trust [4], or trust and link quality [5]. We identify a need for a routing framework that can be easily extended to support any metric.

In this paper, we propose a trust and energy-aware, location-based routing protocol called Trust and Energy-aware Routing (TER) protocol. TER uses trust values, energy levels and location information in order to determine the best paths towards a destination. The protocol achieves balancing of traffic load and energy, and generates trustworthy paths when taking into consideration all proposed metrics. Other metrics can be easily integrated in the protocol.

The protocol relies on the trust values provided by the Adaptive Trust Management Protocol (ATMP), which computes trust based on intrusion detection techniques [6]. However, TER can also use trust and reputation data provided by other trust management mechanisms.

The rest of the paper is structured as follows: Section II presents related work, Section III describes the protocol design, Section IV includes implementation details, Section V presents the experimental evaluation, and Section VI discusses conclusions and future work.

II. RELATED WORK

Based on the network structure, routing in Wireless Sensor Networks can be classified in flat-based, hierarchical-based and location-based routing [7]. Based on protocol operation, routing protocols can be classified in multi-path based, query-based, negotiation-based, QoS-based and coherent-based routing protocols.

Location-based routing protocols compute routing paths based on the location of nodes. Well known location-based protocols are: Geographic Adaptive Fidelity (GAF) [2] and Geographic and Energy Aware Routing (GEAR) [3].

Geographic Adaptive Fidelity (GAF) is an energy-aware, location-based routing algorithm [2]. Location information is used by each node to associate itself to a virtual grid. Nodes in the same grid square are equivalent in regard to packet forwarding and take turns in sleeping and being awake in order to load balance energy consumption. GAF relies on an underlying ad hoc routing protocol.

Geographic and Energy Aware Routing (GEAR) is an energy-aware and location based routing protocol [3]. The protocol selects the neighbor using an energy-aware and geographically informed algorithm to forward the packet towards the target region. Then, it uses a recursive geographical forwarding technique for disseminating the packet in the target region.

Two relevant routing protocols, which take into consideration trust values when determining the path to the

 ISBN: 978-1-61208-203-5

destination are elaborated in TRAP [4] and Zahariadis et al. [5].

TRAP is a trust-aware routing protocol, which represents a component of μRACER routing solution for Wireless Sensor Networks [4]. Each node considers the communication context when choosing the next hop. The communication context includes the past behavior of neighbor nodes and the quality of the links between the local node and the neighbors.

Zahariadis et al. propose the integration of a trust model with a location-based routing protocol [5]. A metric is computed using the distance of the neighbor node to the destination and the trust in the neighbor node. Therefore, the metric is maximized for trustworthy neighbors closer to the destination.

Our protocol is a location-based routing protocol, because it uses the location of neighbor nodes for determining the best path towards the destination. However, TER also considers trust and energy when determining the best next hop. In addition, the protocol uses trust values to determine whether to forward packets from specific nodes.

III. PROTOCOL DESIGN

In Wireless Sensor Networks, most network traffic is upstream traffic, with a many-to-one communication pattern because all packets must reach the base station. In this paper, we develop a method for performing trustworthy routing of upstream traffic.

Trust and Energy-aware Routing (TER) is a trust and energy-aware, location-based routing protocol for Wireless Sensor Networks. The trust values are obtained from Adaptive Trust Management Protocol (ATMP), which computes them based on intrusion detection techniques [6]. We use an extended version of ATMP, which delivers energy and location data along with the trust associations.

TER includes two phases: setup and forwarding. In the first phase, the best next hop towards the base station is selected by taking into consideration several factors, such as trust, energy and location. In the second phase, the packets generated by trustworthy nodes are forwarded using the selected next hop.

A. Assumptions and Notations

A WSN can be represented as a graph, like in Formulas 1, 2 and 3, where N_i are vertices which represent nodes in the sensor network and $\{N_i, N_j\}$ are edges which represent that two sensor nodes can communicate with each other directly.

$$WSN = (V, E) \quad (1)$$

$$V = \cup N_i \quad (2)$$

$$E = \cup \{N_i, N_j\} \quad (3)$$

The set of neighbors of a node is represented in Formula 4, where N_i is the local node and N_j is a neighbor node.

$$NB(N_i) = \{\cup N_j \mid N_j \in V \wedge \{N_i, N_j\} \in E\} \quad (4)$$

The sensor network may be placed in a harsh environment and operate unattended. An attacker may have physical access to the nodes and can compromise them.

We assume that each node knows its location and how much energy it has consumed at any moment. The localization algorithm or technology used for obtaining the location is out of scope for this paper.

We also assume that the Base Station (BS) has a fixed location. Each node knows the location of the BS. This information is distributed to all sensor nodes, during network initialization, along with the shared keys.

The TER assumes that ATMP is extended to send energy and location information along with the trust associations. Therefore, ATMP periodically sends update packets containing the trust associations, the consumed energy and location of the local node (the node sending the updates).

The trust associations (TA) are represented in Formula 5. It includes associations between the neighbors of the local node (n_i) and direct trust values (T_i) [6].

$$TA = [(n_1, T_1), (n_2, T_2), ..., (n_p, T_p)] \quad (5)$$

The update packet (UP) is represented in Formula 6, where E_l is the energy consumed by the local node and (x_l and y_l) are the coordinates of the local node.

$$UP = [TA, E_l, (x_l, y_l)] \quad (6)$$

ATMP takes the trust associations received from multiple neighbors and computes a final trust value. This value has a historical component (T_{old}), a direct (T_d) and an indirect component (T_i), as in Formula 7. The weights are allocated in regard to Formula 8. The final trust (T_{new}) is used in TER when computing the cost.

$$T_{new} = \alpha T_{old} + \beta T_d + \sum_{i=1}^{p} \gamma_i T_i \quad (7)$$

$$\alpha + \beta + \sum_{i=1}^{p} \gamma_i = 1 \quad (8)$$

A node is considered suspicious, if it has a trust value lower than a certain limit (SL), as in Formula 9.

$$Suspicious(N) = \begin{cases} 1 & if\ T(N) \geq SL \\ 0 & if\ T(N) < SL \end{cases} \quad (9)$$

The update packets are authenticated using a broadcast authentication mechanism such as μTESLA [8] in order to prevent malicious updates.

We assume that the parameters of TER and ATMP (weights, limits) can be modified during run-time through generic reconfiguration mechanisms.

The Setup Phase if performed periodically in order to update the costs. The period depends on the number of nodes, topology, mobility, application and security requirements. A large, dense network with mobile nodes, or a network exposed to threats should execute the Setup phase more often.

 ISBN: 978-1-61208-203-5

B. Setup Phase

In the Setup Phase, each node computes a cost for each of its neighbors. The neighbor with the lowest cost is subsequently chosen as the next hop on the route to the BS.

The cost takes into consideration the trust value provided by ATMP, the energy level available on the neighbor node, the distance from the local node to the neighbor node, and the distance from the neighbor node to the base station.

The cost for a neighbor node N is computed using Formula 10, where DT is the degree of distrust in the neighbor node N normalized by the largest distrust among all neighbors, E is the consumed energy of node N normalized by the largest consumed energy among all neighbors, DN represents the distance from the local node to node N normalized by the largest distance, DB represents the distance between N and the BS, normalized by the largest distance, and weights are allocated in regard to Formula 11.

$$C(N) = \alpha DT(N) + \beta E(N) + \gamma DN(N) + \delta DB(N) \quad (10)$$

$$\alpha + \beta + \gamma + \delta = 1 \quad (11)$$

The distrust (dt) is computed from the trust value generated by the ATMP (Formula 7) in regard to a specific neighbor, using Formula 12. In this formula, T signifies the trust value and MaxTrust is the maximum value for the trust parameter. The normalized value of distrust (DT) is computed and used in Formula 10.

$$dt(N) = MaxTrust - T(N) \quad (12)$$

The distance to a neighbor (dn) is computed using the coordinates of the local node and the ones of the neighbor node, as in Formula 13, where x_L and y_L are the coordinates of the local node, x_N and y_N are the coordinates of the neighbor node N. The distance is normalized (DN) and used when computing the cost.

$$dn(N) = \sqrt{(x_L - x_N)^2 + (y_L - y_N)^2} \quad (13)$$

In the same manner, the distance from the neighbor node and the BS (db) are computed, using Formula 14, where x_B and y_B are the coordinated of the BS, x_N and y_N are the coordinates of the neighbor node N. The normalized value (DB) is used when computing the cost.

$$db(N) = \sqrt{(x_N - x_B)^2 + (y_N - y_B)^2} \quad (14)$$

In the Setup Phase, the node computes the cost for each neighbor and chooses the neighbor with the lowest cost as next hop towards the BS. Formula 15 represents the next hop, where N_j is the neighbor node with the minimum cost, and nb is the number of neighbors.

$$NH(N_i) = \{N_j \mid N_j \in NB(N_i) \wedge C(N_j) = \min\{C(N_1), C(N_2), .., C(N_{nb})\}\} \quad (15)$$

C. Forwarding Phase

In the Forwarding Phase, the node receives packets and forwards them towards the base station only if they are trustworthy. The trustworthiness of a packet is determined using Formula 16, where T is the trust in source node SN, TL is the minimum allowable trust limit and MAC is the Message Authentication Node.

$$Trustworthiness(P) = \begin{cases} 1 & if\ T(SN) \geq TL \wedge MAC(P)\ is\ valid \\ 0 & if\ T(SN) < TL \vee MAC(P)\ is\ invalid \end{cases} \quad (16)$$

If the packet cannot be authenticated (MAC) or if the source node has a trust value lower than the trust limit (TL), the packet is considered untrustworthy.

D. Design considerations

Most applications that use WSNs do not require reliable delivery. The use of acknowledgement considerably increases energy consumption. Therefore, we do not include an acknowledgement mechanism in TER. However, if the application does not tolerate packet loss, acknowledgements are easy to integrate with our protocol.

Duplicate detection is necessary in the case of routing loops. However, in order to detect duplicates, information about each packet has to be stored on the nodes. This has a considerable impact on memory usage. If the application requires duplicate detection, TER can be easily extended to support such feature.

IV. IMPLEMENTATION

The protocol has been developed in TinyOS [9], within a layer in the communication stack, between the Active Message and the Application layers. A nesC [10] component has been used for implementing the two phases of TER.

Because TinyOS is an event-based operating system, code is executed only when an event takes place. We have three types of events in TER: receive trust, location and energy data from ATMP, trigger timer, and receive packet. The flow of operations for the three types of events is represented in Figure 1.

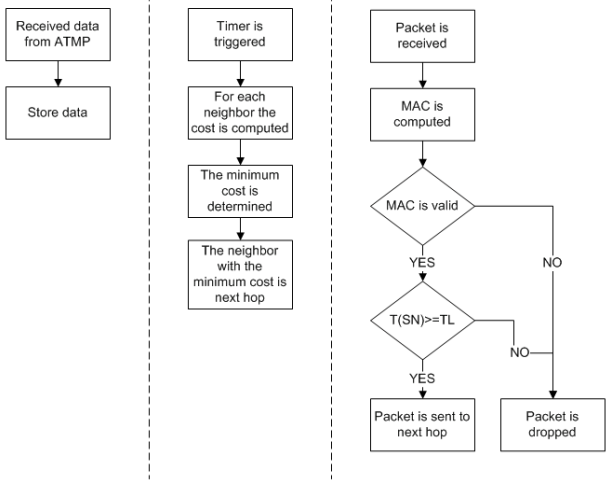


Figure 1. TER workflow

The TER component communicates with the ATMP component through an interface, in order to receive trust, energy and location information regarding the neighbor nodes. The ATMP component sends the data through a nesC event when it has obtained trust, energy and location information. The information is stored by the TER component.

A timer is used for periodically computing the cost using the information received from the ATMP component, according to Formula 10. The component determines the neighbor with the lowest cost and stores the identifier of the neighbor as next hop.

When a packet is received, the first step is to validate the MAC. If the MAC is invalid, the packet is considered untrustworthy and discarded. If the MAC is valid, the trust value for the source node is verified. If the trust value is below a certain accepted limit, the packet is considered untrustworthy and dropped. If the MAC is valid and the trust is above the accepted limit, the packet is forwarded trough the next hop.

V. EXPERIMENTAL EVALUATION

The protocol has been evaluated experimentally using TOSSIM, a simulator for TinyOS [11]. TOSSIM captures the behavior of a large number of nodes at network bit granularity. Therefore, it is a reliable tool for evaluating the behavior of TER enabled nodes in different test cases.

We want to test our protocol in a realistic environment, in order to make sure it operates properly even in harsh conditions. We have therefore used TOSSIM to model an environment with interferences and signal attenuation, which causes a considerable amount of packet loss (30%) specific to harsh environments. The probability of packet loss is increased with the number of hops between source and destination. Therefore, longer paths cause a lower delivery rate.

The test scenario involves a network topology of 10 nodes and the Base Station (Node 0), as in Figure 2. For the analysis, we isolate the flow of packets generated by Node 7 destined to the Base Station.

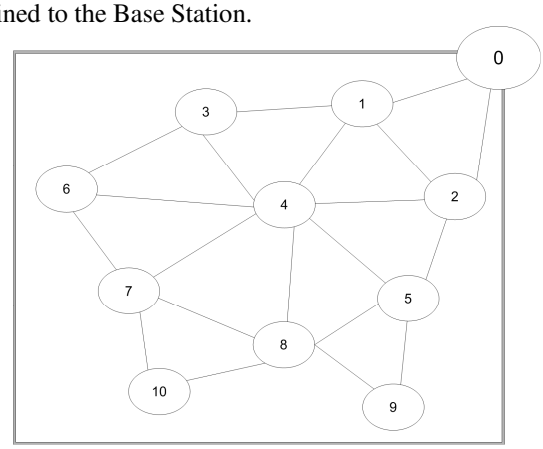


Figure 2. Scenario Topology

We analyze the behavior of TER in different test configurations - with different values for weights α, β, γ and δ. Table 1 includes the analyzed configurations and the values for the considered weights.

TABLE I. TEST CONFIGURATIONS

	α	β	γ	δ
	Trust	Energy	Node-neighbor	Neighbor-BS
Configuration 1	0	0	0	1
Configuration 2	1	0	0	0
Configuration 3	0.3	0.3	0.1	0.3
Configuration 4	0.4	0.3	0.1	0.2
Configuration 5	0.3	0.4	0.1	0.2

We vary the weights from Formula 10 in order to determine the best routing behavior for the proposed scenario. This behavior is evaluated in regard to the number of packets routed through suspicious nodes and to energy consumption.

In Configuration 1, only the distance from the neighbor to the destination is considered, therefore the neighbor which is closer to the destination has the lowest cost. In Configuration 2, only trust is considered: the most trustworthy neighbor has the lowest cost. The first 2 configurations serve as benchmarks in order to determine the influence of a single metric on the packet flow.

The next 3 configurations take in consideration all the proposed metrics and they can be used to determine the most appropriate routing behavior for a specific situation. In Configuration 3, trust, energy and distance to BS have equal weights while the distance to neighbor has a lower weight. In Configuration 4, trust has highest weight, then energy and distance to BS, while the distance to neighbor has the lowest weight. Configuration 5 is similar to Configuration 4 but the energy has the highest weight.

We evaluate the routing behavior of sensor nodes by considering a particular scenario with two suspicious nodes: the trust in Node 4 is 60%, the trust in Node 10 is 40%, the Trust Limit is 50%, and the Suspicious Limit is 80%. This implies that Node 10 sends untrustworthy packets, which will not be forwarded by other nodes.

For each configuration specified in Table 1, we evaluate the routing behavior when delivering a large number of packets generated by Node 7 and destined to the BS. We ran each test 20 times and computed the average values for routed packets and energy consumption, for each considered configuration.

A. *Routed Packets*

A way of evaluating routing behavior is through the number of packets routed by each node. From the results, we can determine which are the most used paths for each configuration, and whether the suspicious nodes are effectively avoided.

The number of routed packets per node, in each configuration, is represented in Figure 3. An average number of 370 packets are sent by the source node 7, as it can be observed in the figure.

For Configuration 1, all packets take the route [7, 4, 1, 0]. This is the best path when taking in consideration the distance between the neighbor and the destination. An

 ISBN: 978-1-61208-203-5

average number of 334.6 packets are delivered through suspicious node 4 (all packets which are not lost on the link between Node 7 and Node 4), but no packet is delivered through suspicious node 10. For this specific topology, the algorithm chooses an efficient path but routes through an untrustworthy node.

In Configuration 2, all packets follow the route [7, 6, 3, 1, 0], the first best path when taking into consideration the trust values. No packet is delivered through the suspicious nodes. For this specific topology, all packets, which are not lost during transmission, reach the base station. This is because the algorithm chooses the first trustworthy next-hop which happens to be placed in the direction of the base station. In other topologies, it is possible that the algorithm does not pick a neighbor in the right direction; in such a case, the paths would be longer and more packets would be lost during transmission.

Because of the greedy algorithm implemented by TER, trust or energy cannot be used as single metric when computing the cost. Therefore, it is better to use these metrics in combination with location.

In Configuration 3, the traffic load is more balanced. The paths that are used for packet delivery are: [7, 8, 5, 2, 0], [7, 6, 4, 1, 0], [7, 6, 3, 1, 0], [7, 8, 9, 5, 2, 0]. The average number of packets delivered through node 4 is 0.65, which is very low. No packets are sent through suspicious node 10. The configuration provides a very good load balancing, as it uses 4 paths to the destination and a very small number of packets are delivered through suspicious nodes.

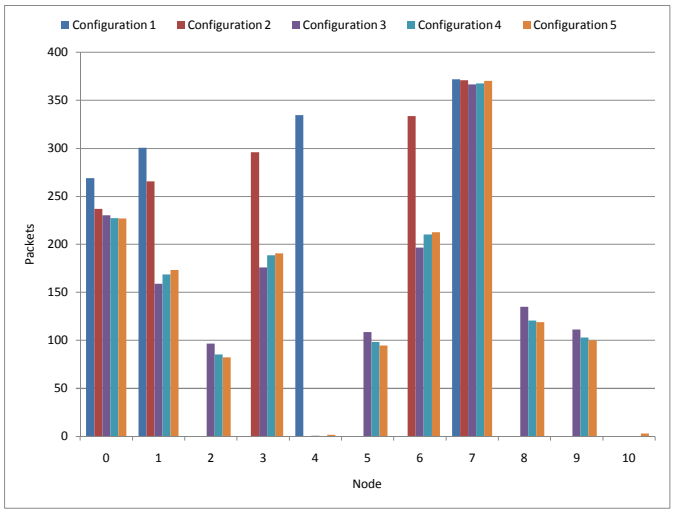


Figure 3. Routed Packets per Node

In Configuration 4, packets are delivered through paths: [7, 8, 5, 2, 0], [7, 6, 3, 1, 0], [7, 8, 9, 5, 2, 0]. No packet is delivered through suspicious nodes 4 and 10. A good load balancing is assured in this configuration and suspicious nodes are avoided.

In Configuration 5, several paths are used for packet delivery: [7, 8, 5, 2, 0], [7, 6, 3, 1, 0], [7, 8, 9, 5, 2, 0]. Some routing loops are generated: [7, 10, 8], [7, 4, 5, 9, 8]. The average number of routed packets through node 4 is 1.65 and through node 10 is 3. The configuration has a good load balancing but it may produce routing loops and a small number of packets are delivered through suspicious nodes.

When analyzing the packets' paths, we determine that Configurations 3 and 4 are the best for this scenario because they have good load balancing, do not create routing loops, and avoid suspicious nodes.

B. *Energy consumption*

Another way of evaluating routing behavior is the energy consumed while routing data packets. We wish to determine whether energy consumption is well balanced between the nodes. The energy metric has an important role in balancing consumption. Without the energy metric, the packets would take the same path and deplete the energy of the nodes on that path.

We evaluate the energy consumption necessary for routing 300 packets generated by Node 7 and destined to the BS. The energy consumed with routing data packets towards the destination, on every node, in each configuration, is represented in Figure 4. The values are represented in Joules. A sensor node has two alkaline AA batteries, each with 9360 J energy. Most energy is consumed with sending and receiving packets. Amiri determined experimentally that a byte sent or received by CC2420 radio consumes 0.12mJ [12].

From Figure 4, we can determine whether energy consumption is balanced between the nodes and if suspicious nodes have been avoided.

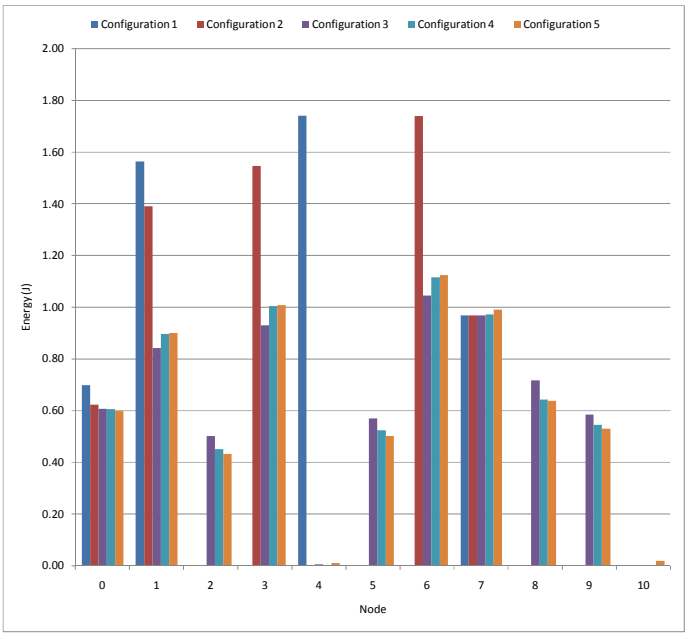


Figure 4. Energy Consumption per Node

In Configuration 1, the most energy is consumed on the suspicious node 4. Energy consumption is not well balanced, as nodes 2, 3, 5, 6, 8 and 9 have no energy consumption due

to packet delivery. Node 7 and node 0 have lower energy consumption than nodes 4 and 1 because they either only transmit or only receive data packets. The energy consumption on nodes 4 and 1 doubles because they transmit and receive packets. Overall, the configuration does not have a balanced energy consumption and routes through suspicious nodes.

In Configuration 2, the suspicious node is avoided, but energy consumption is still not so well balanced, because nodes 2, 5, 8 and 9 are not delivering any packets. The energy consumption drops from 1.74 J on node 4, to 1.55 J on node 3 and to 1.39 J on node 1 because of packet loss. Packets are lost during transmission, so less packets are routed by the subsequent nodes.

In Configurations 3 and 4, energy consumption is well balanced in the network and there is no energy consumption on the suspicious nodes. Configuration 5 is also well balanced and has low energy consumption on the suspicious nodes. These 3 configurations are the best from the point of view of balancing energy consumption due to data packet delivery.

The total energy, consumed on all nodes while delivering 300 data packets generated by Node 7, is represented in Figure 5. The least energy is consumed in Configuration 1 because the protocol determines the shortest path to the destination. Similar energy consumptions have been determined for configurations 3, 4, and 5, which try to determine the shortest path while avoiding suspicious nodes and balancing energy consumption. Configuration 2 has lower energy consumption but it is not well balanced throughout the network.

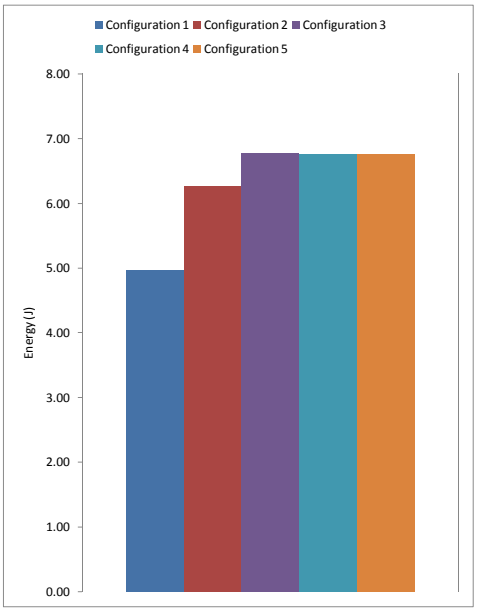


Figure 5. Total Energy Consumption

The energy metric imposes an energy cost but at the same time it allows for a good balancing of energy consumption (see Figure 4), which is an important aspect for Wireless Sensor Networks. The energy metric is particularly important if there is no redundancy among nodes concerning the transmitted information, and therefore we aim to avoid the energy depletion of the nodes which may be preferred by the routing protocol due to their position.

C. *Discussion*

Configuration 1 does not take into consideration neither energy nor trust, including only the distance from the neighbor to the base station. Therefore, it may route packets through nodes which are untrustworthy or have low energy. It only guarantees that it chooses the shortest path towards the destination, as it was observed experimentally. The shortest path consumes the least total energy on sensor nodes but it does not avoid nodes with low power. If the nodes are not mobile, the path is used until some of the nodes die and another path has to be chosen. On the long term, this strategy may determine the partitioning the network. This configuration does not provide load balancing of traffic, is not trustworthy and does not have a balanced energy consumption.

Although Configuration 2 generates trustworthy paths, these paths can be long and inefficient in some cases because the algorithm does not take location into consideration. The only guarantee is that it chooses trustworthy paths. If nodes are not mobile and if the trust values do not change, the algorithm chooses the same path and it consumes all nodes' energy on the path. It does not provide load balancing, it does not guarantee that an efficient path is chosen, and it does not balance energy consumption.

Configuration 3 has a very good load balancing of network traffic, delivers a small number of packets through suspicious nodes and balances energy consumption.

Configuration 4 performs load balancing of network traffic, selects trustworthy and short paths, and balances energy consumption on sensor nodes. Trust has a greater weight and this explains the minimum amount of packets routed through suspicious nodes.

Configuration 5 performs load balancing for network traffic, balances energy consumption, but routes through suspicious nodes, and generates routing loops.

The last three configurations have similar total energy consumption, provide load balancing of traffic, balancing of energy consumption, and they avoid suspicious nodes. From these configurations, Configuration 3 is preferable insofar it has the best load balancing of network traffic and Configuration 4 is preferable insofar it has the minimum number of packets delivered through suspicious nodes.

VI. CONCLUSION AND FUTURE WORK

Wireless Sensor Networks that are used for deploying critical applications such as military surveillance or medical monitoring should provide a high level of security and trustworthiness. Therefore, routing protocols for WSNs should to be designed with security in mind, taking into account multiple metrics that support network availability.

We developed Trust and Energy-aware Routing protocol, which is a location-based, trust and energy-aware routing protocol for sensor networks. The protocol is based on the

 ISBN: 978-1-61208-203-5

Adaptive Trust Management Protocol, which computes trust values based on node behavior.

The protocol uses several metrics: trust values, energy levels, the distance between the local and the neighbor node and the distance between the neighbor node and the destination. These metrics may have different weights when computing the cost of routing a packet through a specific neighbor. The cost is computed based on the metrics and their weights. The neighbor with the lowest cost is chosen as next hop towards the base station.

Trust and Energy-aware Routing protocol has two phases: the Setup and the Forwarding phase. In the Setup phase, the next hop is determined, and in the Forwarding phase, the packets generated by a trustworthy source are forwarded and the others are dropped.

We have implemented the protocol in TinyOS and we have evaluated it experimentally using TOSSIM, comparing 5 protocol configurations for the same scenario. Each configuration has different weights for the trust, energy and distance metrics. For each configuration, the routing behavior has been examined in regard to the paths and packets routed through each node, the consumed energy, and the effectiveness of packet delivery.

Traffic load and energy balancing are very important in Wireless Sensor Networks. In relation to other routing protocols, TER achieves a good balancing of load and energy and generates trustworthy paths, when taking into consideration all proposed metrics: trust, energy and distance.

As future work, we plan to extend the protocol to include other metrics, such as link quality, and to support adaptive weights, allowing, for example, the weight for energy to increase over time. Other extensions we want to implement are duplicate detection and acknowledgements.

We also want to integrate our protocol with another trust mechanism. In addition, we wish to evaluate the protocol in a larger, real-world network.

ACKNOWLEDGEMENTS

This work has been partially funded by the European Commission under grant agreement FP7-ICT-258280 TWISNet project, partially by Google Inc., University Relations and partially by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/88/1.5/S/60203.

REFERENCES

[1] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," Computer Networks, 2008, vol. 52, no. 12, pp. 2292-2330.

[2] Y. Xu, J. Heidemann, and D. Estrin, "Geography-informed Energy Conservation for Ad Hoc Routing," in ACM/IEEE International Conference on Mobile Computing and Networking, 2001, pp. 70-84.

[3] Y. Yu, D. Estrin, and R. Govindan, "Geographical and Energy-Aware Routing: A Recursive Data Dissemination Protocol for Wireless Sensor Networks," 2001.

[4] A. Rezgui and M. Eltoweissy, "µRACER: A Reliable Adaptive Service-Driven Efficient Routing Protocol Suite for Sensor-Actuator Networks," IEEE Transactions on Parallel and Distributed Systems, 2009, vol. 20, no. 5, pp. 607-622.

[5] T. Zahariadis, P. Trakadas, H. Leligou, P. Karkazis, and S. Voliotis, "Implementing a Trust-Aware Routing Protocol in Wireless Sensor Nodes," Developments in E-systems Engineering, 2010, pp. 47-52.

[6] L. Gheorghe, R. Rughiniş, R. Deaconescu, and N. Ţăpuş, "Adaptive Trust Management Protocol Based on Fault Detection for Wireless Sensor Networks," in The Second International Conferences on Advanced Service Computing, 2010, pp. 216-221.

[7] J. Al-Karaki and A. Kamal, "Routing techniques in wireless sensor networks: a survey," IEEE wireless communications, 2004, pp. 1-37.

[8] M. Luk, A. Perrig, and B. Whillock, "Seven cardinal properties of sensor network broadcast authentication," in Proceedings of the fourth ACM workshop on Security of ad hoc and sensor networks - SASN '06, 2006, pp. 1-10.

[9] P. Levis et al., "TinyOS: An operating system for sensor networks," Ambient Intelligence, 2004, pp. 115-148.

[10] D. Gay, P. Levis, R. Von Behren, M. Welsh, E. Brewer, and D. Culler, "The nesC language: A holistic approach to networked embedded systems," in Proceedings of the ACM SIGPLAN 2003 conference on Programming language design and implementation, 2003, vol. 35, no. 11, pp. 1-11.

[11] P. Levis, N. Lee, M. Welsh, and D. Culler, "TOSSIM: accurate and scalable simulation of entire TinyOS applications," in Proceedings of the first international conference on Embedded networked sensor systems - SenSys '03, 2003, pp. 126-137.

[12] M. Amiri, "Wireless sensor networks: Evaluation of power consumption and lifetime bounds," LAP LAMBERT Academic Publishing, 2011, pp. 1-60.

 ISBN: 978-1-61208-203-5

New Traffic Message Delivery Algorithm for a Novel VANET Architecture

Yueyue Li
School of Science and Technology
Nottingham Trent University
Clifton Lane, Nottingham, NG11 8NS, UK
yueyue.li@ntu.ac.uk

Evtim Peytchev
School of Science and Technology
Nottingham Trent University
Clifton Lane, Nottingham, NG11 8NS, UK
evtim.peytchev@ntu.ac.uk

Abstract—**Traffic problems in the field of Intelligent Transport System (ITS) have always been an attraction in the researchers' eyes all over the world. To reduce traffic congestions, to save travel time, to decrease traffic accidents and to provide demanding information exchanges have become challenges of today and the future. Current research works focus on applying Car-to-Car (C2C) and Car-to-Infrastructure (C2I) approaches in infrastructure-less and flexible ad hoc networks environment. The routing problem has always been one of the most difficult problems in such dynamic environment network. This paper presents a novel, designed for routing purposes, traffic routing algorithm (TMDA) for a novel VANET architecture. The algorithm with the inclusion of urban traffic related routing information has been designed to be deployed in vehicles, e.g., cars and buses and aims to provide proper strategies for the utilization of travel information available in many of the vehicles traversing urban networks. The research investigates and compares communication performance of the communication system under TMDA and the other existing ad-hoc routing protocol (e.g., Ad hoc On-Demand Distance Vector) by a set of experiments with the NS-2 simulator. According to simulation-based performance evaluation, the proposed algorithm, TMDA, provides higher efficiency and reliability than a popular used broadcasting method for data dissemination.**

Keywords-ITS; C2C/C2I; ad hoc network; VANET; routing algorithm; NS-2 simulation

I. INTRODUCTION

In recent years, much more projects emerge in the field of Intelligent Transport System (ITS) because of the increasing traffic problems, such as traffic jam and fast accident notifications etc. Fast and reliable real-time traffic information is irreplaceable tool to build safe and efficient traffic environment. To achieve this goal, traffic objects should cooperate with each other by using Car-to-Infrastructure (C2I) and Car-to-Car (C2C) communication approaches, as the communication of information is the biggest unutilised fully factor in ITS for reducing traffic congestions, saving travel time, decreasing traffic accident, improving air pollutions, lowing energy consumption and also providing demanding information during travels.

Typical examples adding weight to this concept are C2X communications investigated in the following projects of the 6th EU Framework Programme for Research and Technological Development [1]: 60 million EU CVIS (Cooperative Vehicle-Infrastructure Systems) Integrated Project [2], targeting mobile traffic participants to provide wise interactions between mobiles and transport infrastructures for road safety; COOPER (CO-OPerative SystEms for Intelligent Road) project [3], aiming at cooperative traffic management by exchanging real-time traffic information among travellers and fixed roadside system to finally enhance road safety on motorways; and SAFESPOT integrated project [4], cooperating intelligent information exchanges between vehicles and roadside units to realize safe and efficient transportations. These projects attempt to integrate C2C and C2I applications while existing outcomes show that the focal point is C2I solutions, by utilizing the supports of roadside units (RSU), access points (AP) and cellular base-stations etc.

While the C2I architectures have been well developed nowadays, further problems about the cost of infrastructure deployment, the speed of connections and the volume of data are considered. Hence, more and more research work and projects pay attention to ad hoc networks, which are self-organized, dynamical and flexible for solving certain urgent social problems, e.g., emergency services and traffic information exchanges, etc. [5] with co-operations of other practical technologies.

In this paper, novel Vehicle Ad-hoc Network (VANET) architecture for city traffic communications is introduced. This framework will create an opportunity for investigation of the benefits of car-based acquisition and dissemination of traffic information as well as generation and distributed implementation of traffic control. For routing purposes, the system applies a new Traffic Message Delivery Algorithm (TMDA). The defining novelty in this algorithm is the presence and utilization of travel route information available in many of the vehicles presenting in the traffic e.g., all buses, cars using Sat-Nav devices etc.

Compared with real test-beds [6][7][8], simulations can save large expenses to construct a model and allow components to execute repeatable tests in diverse targeting scenarios. This paper discusses essential simulation issues via NS-2 and displays results for investigations of the new routing algorithm in the proposed VANET architecture.

 ISBN: 978-1-61208-203-5

The paper is structured as follows. Next section processes literature reviews on broadcasting techniques and introduces the new ideas about essential information being included in the transmission messages. Then newly VANET architecture with a proposed message delivery algorithm TMDA is introduced in details. There are a set of simulation experiments exhibited to evaluate communication performances with the innovative routing protocol. Finally, we conclude results and give a future vision.

II. RELATED WORK

Presently, a plethora of routing protocols is designed to adapt flexible and dynamic ad hoc networks. This paper will only concentrate on those studies being directly related to the proposed techniques and protocols.

A. *Broadcasting in VANET*

Broadcasting is a basic method used in ad hoc networks. The simplest and earliest broadcasting technique is flooding methods, as described in [9][10][11]. Each mobile node, which receives the packet for the first time, periodically broadcasts or rebroadcasts the packet to all neighbours; otherwise, the receiver will discard the packet due to redundant operations. Ho et al. [12] state that a simple flooding method provides minimal state and high reliability, particularly being suitable for highly mobility networks, such as MANET and VANET.

The main problem of the simple flooding, also known as blind flooding [13], is the high amount of redundant broadcasting messages. This is referred as broadcast storm. To solve the problem, a few of solutions have been proposed. For example, a probability-based method from [14] assumes that nodes rebroadcast the received packet depending upon the predetermined probability. If the probability reaches 100%, the scheme is identical to be pure flooding. Additionally, an IEEE802.11-based protocol named urban multi-hop broadcast (UMB) is designed in [15] to minimize the broadcast storm by allowing the farthest vehicles to receive and forward data and inform other nodes between original senders and itself. Meanwhile, it uses acknowledgment messages (ACK) to guarantee high reliability of packet delivery.

As Ros et al. [16] presented, uneven distributions and speeds of vehicles are particular characteristics in VANET networks. Due to these reasons, VANET has to deal with high number of disconnections which may impact on message exchanges. U. Lee et al. [17] introduced periodically broadcasting methods to neighbours. In this case, one-hop neighbours will be able to disperse the message via their mobility to more hops of retransmissions. Moreover, Kitani et al. [22] present a concept of ‘message ferrying' in Inter-vehicle communications, introducing 'bus' as the ferry rather than 'car'. It proposes to improve efficiency of information sharing in sparse areas depending on buses which have regular routes and could collect more traffic information.

In this paper, our new algorithm attempts to improve communication performance by using strategic broadcasting mechanism with the inclusion of traffic route information in the algorithm.

B. *The inclusion of essential information*

In the traffic area, diverse and changeable communication demands and traffic problems can occur at any time. For these reasons, maximum and optimum information are expected to be included in communication protocols by many research and projects. Although there has not been any comprehensive and popular message delivery algorithm meeting the requirements yet, some researchers have proposed algorithms with the inclusion of particular traffic information, for example, the inclusion of the acknowledgments into the periodic beacons for high reliability [16] and the inclusion of vehicles' status and surrounding information in [18], etc.

So far, on the basis of studies in existing literatures, the concept of the inclusion of traffic route information has not been proposed and implemented. Certainly, many projects assume electronic devices such as GPS are installed in most of cars and mobile terminals. Hence, those devices could provide route information to car drivers or other traffic participants. However, this information cannot be easily shared with others unless they are included into the message routing protocols. For the proposed purpose, this research introduces designing a new message delivery algorithm with the inclusion of traffic route information based on a novel MANET architecture.

III. THE NOVEL MESSAGE DELIVERY ALGORITHM - TMDA

A. *The proposed VANET architecture*

Wu [6] introduced a VANET architecture that, based on the background of Car-to-Car/Car-to-Infrastructure communications, involves spontaneous wireless communications occurring within a group of wireless mobile nodes (Figure 1). The architecture integrates features of traditional ad hoc networking technologies and VANET technologies, being used in standalone mode or cooperative connections to the larger Internet [23].

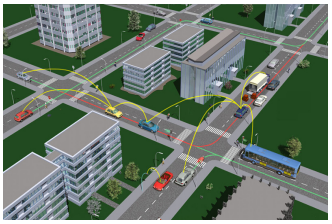


Figure 1. Novel ad hoc wireless mobile network architecture
Ref.: http://www.car-to-car.org/index.php

Being different from traditional ad hoc networks, this communication system utilizes vehicles for routing purposes via the inclusion of traffic route information. It recognises three types of ad hoc nodes - mobile, semi-mobile and static ad hoc nodes. To best exert the functionality of node when communications occur, the system specifies three types of nodes.

Mobile nodes, such as cars, are defined as traditional ad hoc nodes without pre-conceived route with functions of routing and transmitting messages. They could be a major group to request traffic information and fast forward

 ISBN: 978-1-61208-203-5

messages. Indeed, if the car equips high capability electronic devices for message storages, they could carry messages anywhere and exchange to others anytime due to the nature of arbitrariness. However, most of drivers do not accept to spend money on these devices. Hence, car behaviors have to be relatively simplified, e.g., broadcasting only.

Alternatively, bus-nodes, considered as semi-mobile nodes – having predetermined route onto which they are currently traveling, integrate routing, transmitting and gateway altogether to provide a possibility of interconnection among other types of networks, e.g., Internet. Although they could not move everywhere as cars do, they are able to equip powerful devices to offer more communication capabilities than other common vehicles. Typical examples are the energy of transmission, the range of communications, as well as the storage of messages. These are possible to compensate discontinuous delivery occurring between car communications. Moreover, buses and bus-lanes present some particularities in urban scenarios. Most of cities specify lanes for buses priority to guarantee unimpeded travels for the public, even in peak time.

As far as static ad hoc nodes are concerned, they will cooperate with other two types of nodes to provide more reliable and specific information if exchanges of a message between first two kinds of nodes does not meet users' requirements. In this research, static nodes belong to a kind of ad hoc nodes; however, the essence is similar as roadside units. The nodes are expected to provide access for larger scale of information exchanges.

B. TMDA overview

Traffic Message Delivery Algorithm (TMDA) is a novel traffic routing algorithm designed for improving communication performance of a particular VANET network described in Section A. The difference as compared to another routing protocol is that TMDA does not only implement single broadcasting approach, such as the simple flooding, probability-based method, area-based method and neighbourhood-based conception [19], but also adopts intelligent routing strategies by utilizing the pre-existing travel information for message delivery at any given moment. It means that the algorithm with the inclusion of traffic route information will be embedded in each communication mobility node with current advanced information adaptation devices and provide optimization routes to messages between the source and the destination.

TMDA utilizes features of each type of nodes for efficient and reliable traffic communications. For example, it does not only take advantage of arbitrariness of car-nodes, but also exploits the benefits of controllable, scheduled, and predicted bus-nodes; it does not only allow simple broadcasting behaviours of cars, but also make uses of higher capability of bus-nodes for properly storing and forwarding the messages. Furthermore, TMDA is prone to regional message delivery and does not exclude the possibility of Internet access via static nodes to spread messages widely.

C. Algorithm details

TMDA could be divided into two sections: sending and receiving. Procedures are relatively simple for sending the message that nodes carry on periodic broadcasting via IEEE 802.11 within a certain expiry; whereas more considerations occur in terms of receiving a message. Algorithm I is showing the pseudo-code of TMDA in message receiving section.

ALGORITHM I. PSEUDO-CODE OF TMDA IN MESSAGE RECEIVING

```
Event: the message has been received
if msg_id is not in check_list then
    receives the message
else
    discard the message

Event: the message received from NB or S
if R = src then
    discard the message;
else
  if R = dst then
    inform others to stop broadcasting;
  else
    if P_s is on I-Routes then
      if P_r is on I-Routes then
          when T_c = T_d1, farthest nb forward message;
          inform others between <S,R> to stop broadcast;
          message is stored longer in this node R;
      else
        if D_r = D_s then
          when T_c = T_d2, farthest NB forward message;
        else
          when T_c = T_d3, farthest NB forward message;
    else
      if P_r is on I-Routes then
          when T_c = T_d1, farthest nb forward message;
          inform others between<S,R> to stop broadcast;
          message is stored longer in this node R;
      else
          when T_c = T_d1, farthest nb forward message;
```

Actually, above steps implement a selective forwarding mechanism by utilising additional urban traffic related information. The overall aim is to address broadcast storm problems. Two main parts are involved in the mechanism.

One is the idea of I-Route. This is a critical route, e.g., bus lanes, used to determine next operations of nodes. Briefly, if messages reach I-Routes, they will be faster forwarded following the pre-configured directions of the I-Routes; otherwise, they are based on developed broadcasting strategies only. The nodes on I-Route, regardless the real type, are treated as buses. On the basis of I-Route, another concept is about 'farthest node first send' (FNFS). Once a sender delivers a message to all neighbours, the farthest one within the transmission range will deal with the message following

 ISBN: 978-1-61208-203-5

the priority over others. The priority level is set by delays introduced in the following pseudo-code of TMDA. The idea is beneficial to control data collisions to a certain degree.

Message receiving function is divided into two events. From line 1 to 5, when a receiver *R* obtains a message with the id *msg_id*, *R* should firstly check whether it receives a redundant message. Each VANET node has a *check_list* to store received *msg_id*. Thus if the *msg_id* is found in the list, *R* discards the message; otherwise, continues the steps of another event (line 6 to 29).

When *R* receives the message from its neighbours *NB* or source *S*, it needs to make sure that the message dose not loop back. Then if *R* is the destination node, it simply broadcasts back to all neighbours with a stop instruction. Alternately, if *R* is an intermediate node only, steps from line 13 to line 29 are focused on. To judge when to forward the message to neighbours, *r* needs to know *nb*'s or *s*'s position (x, y) and its own position. This helps to check whether they are on *I-Routes* or not. If both of *S* and *R* are on *I-Routes*, then *R* forwards the message at T_{d1} which consists of current_time (*Tc*) and a waiting delay d_1. Within the transmission range, the delay d_1 will be reduced accompanying with the increase of distance between <*S, R*>. That is, the farthest *R* will forward message firstly. Additionally, if *S* is on the *I-Route* but *R* is not, the moving directions of *R* and *S* become important. Same direction of *R* and *S* ($D_r = D_s$) makes the forward occur at T_{d2} while the message is broadcast at T_{d3} for different directions of *R* and *S*. The value of T_{d2} or T_{d3} is different but both consist of a current time *Tc*, a delay according to the distance d_1 and a pre-configured delay d_2 setup by the algorithm. The value order is $T_{d1} < T_{d2} < T_{d3}$.

IV. SIMULATION ISSUES

NS-2 is selected as a well-suitable simulation tool in this paper. It uses Tcl (Tool Command Language) to organize script files for setting up traffic patterns such as scenarios and movements and also communication patterns, e.g., transmission issues.

A. A City scenario

In terms of traffic patterns, the focus at this stage is #-shaped city scenario (Figure 2). Compared with T-shaped patterns in a previous paper [6], this scenario contains more traffic situations.

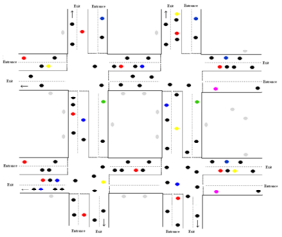


Figure 2. #-shaped traffic pattern

#-Shaped city scenario (Figure 2) – a medium scale network with possible traffic units consists of intersections, horizontal and vertical roads. It can be useful to investigate some issues that whether I-Route areas provide efficient decisions for message delivery; whether different types of nodes work properly to provide high reliability in various densities of networks etc.

Nodes – The term density represents as the number of nodes over the network. This paper presents four dense levels (Figure 3), from very low to high.

Figure 3. Simulation models for example densities of networks

I-Route – This is a term for a set of special routes integrated in our established ad hoc wireless mobile communication system. On I-Routes, message transmissions obey special strategies and they are expected to support for performance improvements. Therefore, I-Routes should have a capability to centralize more mobile nodes so that strategies can be best used. According to features of buses mentioned in previous sections, I-Routes are pre-set to be bus lanes in this paper. This point will be further investigated and validated. Current simulation models adopt the following I-Route patterns, drawn as two lines with arrows in Figure 4. Future more I-Routes could be identified by buses or be pre-configured by control centres due to the different purposes.

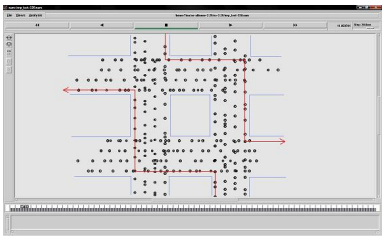


Figure 4. Simulation model with I-Routes

B. Transmissions

Following points, such as transmission range and nodes distance etc. are essentially to impact on the design of simulation models.

Distance – The distance of a node-pair varies because of simulation initializations and node densities. In our designs, the nodes are distributed following the shape of urban lanes and the distance between two nodes is chosen randomly but between 10 to 150 meters. Actually, the value is decided particularly in this research because of real traffic considerations. Meanwhile, the transmission range is set as the same value.

Speed – Regarding to the real world conditions, the speed of vehicles should be different according to transportation conditions, such as the traffic flows, the speed of front nodes and the traffic rules etc. Therefore, the speed of nodes is assigned randomly when nodes are running with different directions.

Time - Total simulation time for above models is set to 300 seconds. Message sending time is randomly chosen by

 ISBN: 978-1-61208-203-5

NS-2 within the total simulation time. We assume that the maximum expire time of message is no more than 60 seconds for non-emergency messages.

Message – Message contains three elements: message size, message id and other information, such as source node, destination node, current sender, the position of senders, the speed of senders, the direction of senders, the message expiry and current timestamp. It assumes that only one message is transmitted between a pair of nodes each time and the minimum number of message over the network at the time is 1 while the maximum value is 10 in this paper.

V. RESULTS EVALUATION AND ANALYSIS

A. Network communication performance metrics

End-to-End Delay Time (EDT) - It refers to the duration of a message sent from source to destination over the network [21]. Note that the equation (1) is used for calculating single-pair of nodes' delay (EDT). Te stands for the end time of a packet delivery and T_0 means the start time; (2) solves the average delays (ΔEDT) by using the sum of single delays (Σ(EDT)) and the number of tests (n).

$$EDT = Te - T_0 \qquad (1)$$

$$\Delta EDT = \Sigma (EDT) / n \qquad (2)$$

The acceptable maximum delay time is limited as 60 seconds for non-emergency messages. If the delay time is over 1 minute, then packet loss is recorded.

Message Delivery Ratio (MDR) – It represents a ratio of successful message deliveries. In equation (3), a single rate is calculated using the number of successful receives (n_r) and the number of original sends (n_s). The final evaluation of this paper will follow the results obtained via equation (4) which shows the average value of the testing delivery ratios.

$$MDR = (n_r / n_s) * 100 \qquad (3)$$

$$\Delta MDR = \Sigma (MDR) / n \qquad (4)$$

B. The Comparison of routing protocols

AODV – Wireless Ad hoc On-Demand Distance Vector (AODV) routing protocol concerns on mobile ad hoc networks (e.g., MANETs) nowadays. It is a reactive routing protocol which creates a route for nodes only when they demand it, being one of common broadcasting routing protocols used currently for both unicast and multicast routing. The serious problem is the broadcasting storm, which attempts to be avoided and reduced in the proposed routing protocol TMDA.

TMDA – Traffic Message Delivery Algorithm delivers messages depending on the concept of pre-configured routes (I-Routes) in the city scenarios. On the basis of general broadcasting methods, TMDA reduces broadcast storms via selective forwarding mechanism, coupled with geographic information.

Table I shows advantages and disadvantages of AODV, which have been proposed and validated for long years. Following that, the anticipated features of TMDA, being given in advance, will be investigated by simulation results in later sections.

TABLE I. COMPARISONS OF ROUTING PROTOCOLS

Routing Protocol	Advantages	Disadvantages
AODV [20]	1) On-demand 2) Destination sequence numbers to find latest route 3) Small control and data packet requires few bandwidth 4) Link broken response fast 5) High reliability in medium and large networks	1) stale entries 2) Multiple RREP packets to a single RREQ packet causes big control overhead 3) Battery and bandwidth consumptions
TMDA	***Anticipated:*** 1) Simple broadcasting mechanism 2) No network topology maintenance 3) No complex route discovery algorithm 4) I-Routes are set up for controlling packet forwards 5) Reduction of broadcast storm	***Anticipated:*** 1) Bear with a certain delays if nodes are not on pre-configured routes 2) Not good for emergency message exchanges in sparse networks

C. Results in various dense networks

Figures 5, 6, 7 and 8 compare EDT and MDR results by applying Traffic Message Delivery Algorithm (TMDA) and implementing Ad hoc On-Demand Distance Vector (AODV) routing protocol in very low, low, medium and high density of networks separately. There is an assumption in the experiments that the acceptable delivery time for non-emergency message is no more than 60 seconds, and random source-to-destination pairs are allowed to exchange various amount of messages (from 1 to 10) per randomly testing time. The overall aim is to investigate whether TMDA leads to less EDT and higher MDR in various scenarios rather than an another existing routing protocol; how degree the amount of messages impact on communication performance; and how the trend of EDT and MDR changes in different network conditions.

1) High & Medium density

Figure 5 represents the average EDT and Figure 6 shows the trend of MDR in the dense and moderate dense network respectively. According to above line charts, TMDA exhibits smaller EDT from 1 message to 10 messages per testing time, reflecting on the below lines in Figure 5 and higher MDR from the above lines in Figure 6 than those obtained from AODV protocols.

 ISBN: 978-1-61208-203-5

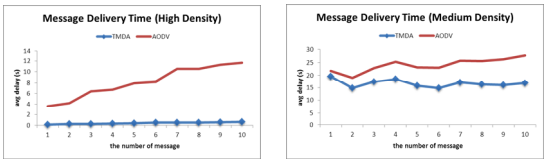


Figure 5. Delays in the high & medium density of networks

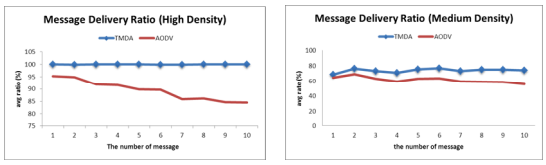


Figure 6. Rates in the high & medium density of networks

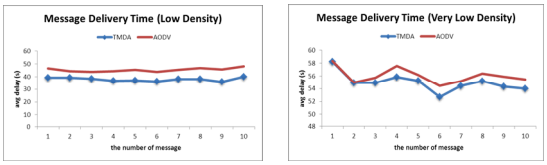


Figure 7. Delays in the low & very low density of networks

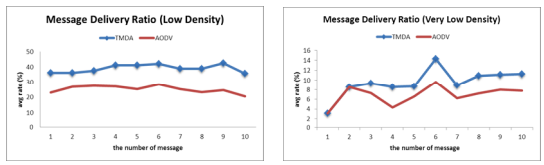


Figure 8. Rates in the low & very low of networks

For AODV, the trend of the average EDT in both networks goes up accompanied with increases of the message number shown in Figure 5; conversely, the ratio displays as decreasing status in Figure 6. Therefore, the number of nodes over the network and the number of transmission messages have significant impacts on the transmission delays and reliability. However, the trends of average EDTs and MDRs are relatively stable when TDMA is used for message deliveries. Particularly in the dense VANET, the average value of EDTs is very small, presenting a distinguished gap between the line of AODV and the line of TMDA. Oppositely, the trend of average MDRs in TMDA keeps in a high level (e.g., 80%-100%) while AODV experiences decreasing values when increasing the message number from 1 to 10.

Compared to the results in moderate density of networks, the results are notably better in the high density network. One of drawbacks inherited from AODV is the broadcast storm which is also considered as a major reason of packet loss. If 10 messages are transmitting over the network, more nodes mean higher possibility to generate data collisions over the network. As introduced earlier in the paper, TMDA adopts delay strategies to reduce broadcast storm and the results prove that the packet loss is relatively less prominent.

Certainly, when the nodes are reduced, both routing algorithms are influenced, reflecting on the increasing delays and the declining packet ratios, e.g., those in medium density of networks. It is understandable that the condition of re-send becomes frequent.

2) Low & Very low density

Figure 7 and Figure 8 display average EDT and MDR in low and very low density networks respectively. TMDA provides better results than those of AODV. For example, EDT lines of TMDA in both networks are lower than those of AODV with smaller average delays. Meanwhile, the above MDR lines which represent higher successful packet deliveries are from TMDA.

Usually, a problem of disconnection seriously occurs in sparse networks. This is because nodes are not enough to forward messages and they are not distributed evenly. The problem causes transmission failures as high possibility of packet loss within an expiry. If the transmission fails within the expiry, AODV provides a sequence of procedures such as packet requesting, replying and repairing etc. to deal with these failures. However, the mechanism suffers more delays because senders should wait reply packets from the destination nodes and then judge if they need to re-send again or stop sending. For TMDA, it allows senders to continually broadcasting the message within the expiry unless they receive a redundant message or they receive an instruction included in the message to stop broadcasting. This approach saves the time for senders to wait the response and also each sender needs not to keep a list to record paths for replying packets.

Moreover, TMDA contains I-Route information. Nodes on the I-Routes are allowed to have longer storage time than nodes on the common lanes. This strategy helps to improve the ratio of message deliveries, particularly in sparse networks. One of cases in the experiments as follows: suppose a source node and a destination node are far from each other and a bus running on the I-Route could pass over each other in a certain time range. AODV allows the bus to re-broadcast the message within T and the distance takes t for the bus to connect with the receiver. Due to T<t, the packet will be dropped. Instead, TMDA allows the bus to extend re-broadcast time to be T1 (T1 > T >= t), then the message could be received. Certainly, in specific cases, the delivery time will be very long by using TMDA, but it could be accepted with a tolerance limit. In our experiments, we set maximum expiry for non-emergency messages to be 60 seconds. That is, any delay time more than 60 seconds will be regarded as final packet loss.

Besides the above features of I-Routes, they could direct message towards assigned directions. If both source and destination nodes are on 'I-Route', the delay could be very small because nodes on 'I-Route' have the high priority of forwarding actions. As in AODV, it lets the message be sent with the same rights of broadcasting requests, replies and forwarding to all one-hop neighbours. Certainly, if the source-to-destination pair is not on the I-Route or not all on

 ISBN: 978-1-61208-203-5

the I-Route, the transmission time could be at least the similar as AODV results. Generally, the average message delivery time, seen in Figure 7, are smaller by using TMDA from 1 message to 10 messages.

VI. CONCLUSION AND FUTURE WORK

This paper presented the comparisons of communication performance by using different routing protocols in a novel VANET architecture. AODV is a published protocol used commonly in ad hoc networks, whereas, TMDA is a newly created algorithm. It not only adopts principles based on existing broadcasting algorithms but also incorporates urban traffic route information into the algorithm, utilizing the concept of 'I-Route' available in vehicles. The aim of these new routing strategies is to alleviate the impact of the problems caused by previous routing protocols and also best service for the particular implementation background. We design a VANET architecture which contains three types of ad hoc communication objects - mobile, semi-mobile and static ones.

So far, investigations indicate that TMDA generally shows better results than the others one in terms of packet delivery time and successful packet delivery ratio in dense, moderate dens, sparse and very sparse networks. The future work will concentrate on applying the algorithm in a real city scenario (e.g., Nottingham city) to further investigate above results of simulations. Meanwhile, static nodes are considered to be integrated into the architecture for collaboration studies.

ACKNOWLEDGMENT

The authors wish to thank all members of the Wireless Communications Simulation Modelling group of Nottingham Trent University, including Dr. Taha Osman, Dr. Richard Germon, Dr. Xiaoqi Ma, Postgraduate Researcher Emad Gamati and all other colleagues.

REFERENCES

[1] European Commission Community Research-CORDIS. The Sixth Framework Programme in brief. 2002

[2] CIVS (Cooperative Vehicle-Infrastructure Systems), [online]. Available: http://www.cvisproject.org [April, 2012]

[3] COOPER (CO-OPerative SystEms for Intelligent Road), [online]. Available: http://www.coopers-ip.eu [April, 2012]

[4] SAFESPOT, [online]. Available: http://www.safespot-eu.org [April, 2012]

[5] J. Wu, Handbook on *theoretical and algorithmic aspect of sensor, ad hoc wireless, and peer-to-peer networks*, United States of America: Auerbach Publications Taylor & Francis Group, 2006.

[6] Y. Li and E. Peytchev, *Novel ad-hoc wireless mobile communication network routing model for location based sensor networks*, in proceedings of International Symposium on LBS & TeleCartography, pp. 383-396, September 2010.

[7] APE: *Ad hoc Protocol Evaluation testbed*, Department of Computer Systems at Uppsala, Sweden.

[8] J. Broch, D.A. Maltz and D.B. Johnson, Quantitative lessons from a full-scale multi-hop wireless ad hoc network testbed, in: Proceedings of the IEEE Wireless Communications and Network Conference 2000 (WCNC 2000), Chicago, IL, pp. 992-997.

[9] T. Camp and B. Williams. Comparison of broadcasting techniques for mobile ad hoc networks. In *Proceedings of The Third ACM International Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC 2002)*, Lausanne, Switzerland, Jun 2002.

[10] A.M. Hanashi, A. Siddique, I. Awan, and M. Woodward, Performance evaluation of dynamic probabilistic broadcasting for flooding in mobile ad hoc networks, Simulation Modelling Practice and Theory, Volume 17, Issue 2, February 2009, Pages 364-375, ISSN 1569-190X, 10.1016/j.simpat.2008.09.012.

[11] J. Jetcheva, Y. Hu, D. Maltz, and D. Johnson. A simple protocol for multicast and broadcast in mobile ad hoc networks. Internet Draft: draft-ietf-manetsimple-mbcast-01.txt, July 2001.

[12] C. Ho, K. Obraczka, G. Tsudik, and K. Viswanath. Flooding for reliable multicast in multi-hop ad hoc networks. In *Proceedings of the International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communication (DIALM)*, pages 64–71, 1999.

[13] I. Stojmenovic and J. Wu, "Broadcasting and Activity Scheduling in Ad Hoc Networks," Mobile Ad Hoc Networking, S. Basagni, M. Conti, S. Giordano, and I. Stojmenovic, eds., pp. 205-229, IEEE Press, 2004.

[14] S. Ni, Y. Tseng, Y. Chen, and J. Sheu. The broadcast storm problem in a mobile ad hoc network. In *Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM)*, pages 151–162, 1999.

[15] L.J. Li, H.F. Liu, Z.Y. Yang, L.J. Ge, and X.Y. Huang, Broadcasting Methods in Vehicular Ad Hoc Networks, *Journal of Software*, Vol.21, No.7, July 2010, pp.1620−1634 (in Chinese with English abstract).

[16] F.J. Ros, P.M. Ruiz, and I. Stojmenovic, "Acknowledgment-Based Broadcast Protocol for Reliable and Efficient Data Dissemination in Vehicular Ad Hoc Networks," *Mobile Computing, IEEE Transactions on* , vol.11, no.1, pp.33-46, Jan. 2012

[17] U. Lee, J. Lee, J. Park, E. Amir, and M. Gerla, "FleaNet: A Virtual Market Place on Vehicular Networks," Proc. Third Ann. Int'l Conf. Mobile and Ubiquitous Systems: Networking and Services, pp. 1-8, July 2006.

[18] J. Fukumoto, Sirokane, Y. Ishikawa, Wada, N. Tomotaka, K. Ohtsuki, H. Okada, Analytic method for real-time traffic problems in VANET by using Contents Oriented Communications, 7th International Conference on ITS, 6-8 June 2007.

[19] N. Karthikeyan, V. Palanisamy, and K. Duraiswamy, *Performance Comparison of Broadcasting methods in Mobile Ad Hoc Network*, International Journal of Future Generation Communication and Networking , Vol. 2, No. 2, June, 2009.

[20] S. Hamma, E. Cizeron, H. Issaka, and J.P. Guedon, 2006. *Performance evaluation of reactive and proactive routing protocol in IEEE 802.11 ad hoc network.* In: Proceedings of ITCom 06; ITCom 06 - next generation and sensor networks,2006-10-02.pp. 638709.

[21] L. Feeney, B. Ahlgren, A. Westerlund, Spontaneous networking: an application-oriented approach to ad hoc networking, IEEE Communications Magazine (2001).

[22] T. Kitani, T. Shinkawa, N. Shibata, K. Yasumoto, M. Ito, and T. Higashino, "Efficient vanet-based traffic information sharing using buses on regular routes," May 2008, pp. 3031–3036.

[23] I. Chlamtac, M. Conti, and J.J.-N. Liu, Mobile ad hoc networking: imperatives and challenges, SCIENCE@DIRECT, 2003.

 ISBN: 978-1-61208-203-5

Evaluating SLAM Approaches for Microsoft Kinect

Corina Kim Schindhelm
Siemens AG – Corporate Research & Technologies
Munich, Germany
Email: corina.schindhelm@siemens.com

***Abstract*—The weak performances of GPS within buildings is the reason for a lot of different approaches for indoor positioning, e.g., by using WiFi or odometry. The current position of a person is crucial, for example, for location based services that increase not only outside of buildings. Navigation systems in subway stations is just one obvious example, where GPS fails to deliver the necessary information. Especially in the field of visual odometry, there are many approaches. But all of them are either based on normal 2d camera systems or on expensive 3d camera systems. In the presented approach, we use a Microsoft Kinect, as these systems are inexpensive and widespread. We evaluate how different state of the art techniques like RANSAC or ICP can be used in combination with the Kinect and how they perform in different indoor scenarios. Our evaluation shows that those techniques can be used for the Kinect but have their shortcomings in different scenarios. For that reason, a hybrid technique was developed which combines those methods using a Kinect specialized ICP weight function. In addition, we use a loop detection algorithm to further optimize the accuracy. Finally, we present our results obtained during tests in three different test environments. This paper presents the result of different SLAM approaches implemented on the Microsoft Kinect in order to calculate trajectories.**

Keywords-slam, kinect, odometry, indoor positioning.

I. INTRODUCTION

Many indoor positioning methods have been researched and some solutions found their way into consumer products. But there are still not many (public) buildings equipped with indoor positioning systems, even though it would add value to many public institutions (e.g., libraries, schools, universities) or other areas without satellite coverage (e.g., subway stations, tunnels). Mostly, indoor positioning solutions have been deployed into companies with sufficient funds to invest in expensive high precision technologies like Ultra Wide Band, since their businesses can directly benefit through use of indoor asset tracking [1].

A different approach to installing expensive indoor positioning solutions, which also often need a lot of calibration and maintenance, is to make use of a method known from the field of robotics called SLAM (simultaneous logging and mapping). The main idea there is to place a mobile robot at an unknown location in an unknown environment and let the robot incrementally build a consistent map of its environment while simultaneously determining its location within this map [2]. There exists a lot of different algorithms and solutions to solve this problem. We were interested in the question whether those approaches can also be applied to humans and everyday devices instead of robots equipped with high-end sensors.

This paper deals with the comparison of two different SLAM methods and a hybrid approach, which are applied to the Microsoft Kinect carried by a human being. We developed an evaluation platform which allows to compare different SLAM algorithms and their performance in different scenarios (test environments).

The reminder of this paper is structured as followed: Section II will introduce SLAM principals and list some reference work in this field. Section III describes the Microsoft Kinect, the concept and the three different test environments. Section V evaluates the implemented algorithms in respect of the test environments and Section VI concludes the paper.

II. FUNDAMENTALS OF SLAM

SLAM is a method usually applied by robots to create a map of the surrounding while at the same time estimate their location. Among the vast number of different SLAM methods the main principal remains the same: At the start there is no map of the environment, hence the position of the robot is the origin of the coordinate system and the measurement at this position is the initial measurement. From then on each subsequent measurement contains already known data and new unknown data. By comparing the current measurement with the data set the robot can find am overlapping, and therefore, calculates its new position. By including the new measured data into the map, the whole environment can be surveyed incrementally. Since the position shift between two measured data sets is not perfect, the map quality decreases over time. Tim Bailey and Hugh Durrant-Whyte offer two tutorials about SLAM, which deal with the SLAM problem and algorithms solving the problem [2][3]. As mentioned before, there exists a vast number of SLAM methods, e.g., algorithms using particle filters , the Extended Kalman Filter or graph based techniques.In this paper we will focus on two different approaches: the first is based on visual key points and the second one is based on point clouds. Further details will follow in Section III. In practice, there is a variety of systems based on SLAM that use different sensor equipment. A SLAM system using INS (inertial navigation systems) was developed by Robertson et al. INS sensors were installed to

 ISBN: 978-1-61208-203-5

pedestrians' feet to obtain 2D maps of large areas based on iterative processing of pedestrian odometry data [4]. A system using an Extended Kalman Filter and laser scanners was developed by Garulli et al. [5]. Multiple robots using landmarks to create independent maps, which have to be combined subsequently were used by Zhou [6]. A systems using cameras and SURF detectors was implemented by Engelhard et al. [7].

III. SLAM WITH KINECT

This section offers hardware data of the Kinect, the SLAM algorithms and the information about the test settings.

A. The Kinect and quality of sensors

The technical components for the Kinect were developed by PrimeSense [8], which also published the open source API OpenNI together with WillowGarage [9] and Side-Kick [10]. PrimeSense patented Light Coding generates depth information based on a infrared laser projector and a monochrome CMOS sensor camera. The resolution of Kinect's depth image is 320 x 240 pixel, which is internally interpolated to the double size of 640 x 480. Objects can be recognized to a distance within the range of 0.8 meter to 6 meter. The horizontal field of view is 57 and the vertical 43 [11]. An additional RGB camera provides 640 x 480 pixel color images. Together with an audio channel, the micro processor offers a synchronized data stream of color, depths and audio information to a rate of 30 Hz [12].

Since SLAM algorithms are based on accurate sensor data we examined the error rate of Kinect's depths information. The test comprised a set of Kinect pictures of a simple wooden board placed parallel to the view of the Kinect. Measurements were taken from different distances. Figure 1 shows the result that with bigger distance the error of raw data grows significantly. A picture taken from four meters distance results in a maximum of 14.2 centimeters deviation, whereas with 80 centimeters distance the maximum deviation is only one centimeter. To reduce the errors which mainly result from signal noise we applied and examined different filters. Exemplary the results of a median filter [13] and a bilateral filter [14] with different parameters are depicted in Figures 1 and 2. The figures show that using filters can help minimizing the deviation.

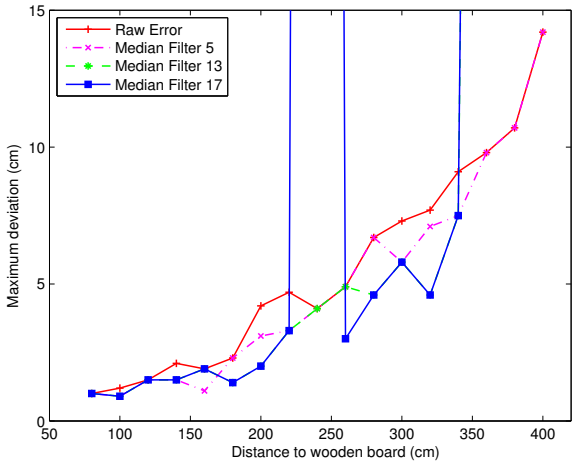


Figure 1. Medianfilter

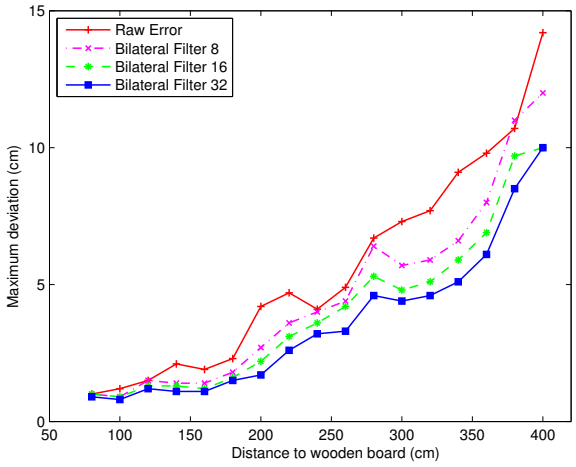


Figure 2. Bilateral filter

B. SLAM Algorithms

Figure 3 gives an overview of the algorithms that are implemented and examined: Visual Keypoints (upper part of Figure 3), Hybrid (middle part of Figure 3) and ICP (lower part of Figure 3).

The SLAM method based on visual key points (see Figure 3 upper part) works as follows: In the first step striking key points have to be detected and categorized (e.g., SURF and Shi Tomasi). The SURF(Speeded Up Robust Feature) method [15] is an enhancement of the SIFT(Scale-invariant feature transform) method [16]. The goals of both is to robustly identify key points among disordered data with a descriptor invariant to uniform scaling, orientation, and partially invariant to distortion and illumination changes. The advantage of the SURF is the higher speed which is achieved for example by replacing the Gaussian filter with a Box filter. Shi and Tomasi detectors are based on Harris and Moravec detectors. The goal of those approaches is to detect corners, whereas a corner is defined as a point with low self similarity. Afterwards, in a second step, homologous key points in two subsequent picture frames must be found. Key points between two pictures found with SIFT/SURF detectors and descriptors can be matched with the minimal Euclidean distance. For key points found with the approach of Shi and Tomasi, the optical flow is applied. In a final, step homologous key point pairs are used to calculate the position

 ISBN: 978-1-61208-203-5

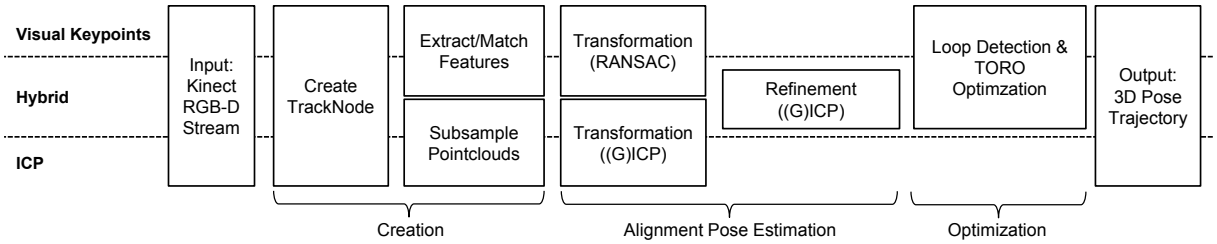


Figure 3. Overview of application flow of visual key point, hybrid and ICP approaches

transformation (RANSAC [17]). The goal of the RANSAC algorithm is to find a suitable model that describes the position transformation best. The algorithm can be described in 4 steps: 1. Select randomly sufficient homologous key point pairs. 2. Define a possible characteristic of the model. 3. Apply this model to all key points of the first picture. The key point pairs fitting this model are defined as inliers. 4. Calculate the quality of the model and decide whether the new model with the number of inliers is better than the current model. If so, the new model is now the best model. This procedure is repeated a prior defined fixed number of times, each time producing either a model which is rejected because too few inlier points were found or a better model with lower error measurement. The RANSAC is very robust to noise and measurement errors and outliers, but the number of iteration has to be limited since it is a non-deterministic approach, which may result in an imprecise or even incorrect model. Finally, the TORO (Tree-based network optimizer) optimization is performed [18]. The resulting graph of RANSAC underlies the general problem of all SLAM methods. The errors in sensor measurement cumulates over time and results in a deviation that also increases over time. In case a position is passed twice, pictures and key points can be recognized and a loop is detected. The goal of TORO is now to minimize errors of the calculated positions, which might have occurred since the time when the position was passed the first time.

The second method (see Figure 3 lower part), the ICP (Iterative Closest Point) Application Flow, uses point clouds as input to calculate position transformations. The Generalized ICP [19] takes two partly overlapping or completely identical point clouds and aligns them until they match. The algorithm works in two steps: Find correspondences between both sets of point clouds and iteratively revises the translation and rotation needed to minimize the distance between the two sets. The correspondences can be weighted either with a Point-to-Point Minimization [20] or a Point-to-Plane Minimization [21].

We also examined a hybrid application flow (see Figure 3 middle part), which works in the beginning like the visual key point application flow, but performs a refinement with the ICP in the Alignment Pose Estimation Phase. In the case not enough homologous key point pairs could be found, the algorithm immediately switches to the ICP calculation, which ensures that even in situations where visual SLAM fails a position can be calculated and gaps in the output graph prevented.

C. Evaluation platform and test environments

The evaluation platform offers several features to ensure consistent and comparable results: All algorithms must have the same input data (Kinect data stream). Hence, the platform offers a record function, where each walking path is stored into an ONI file. Each algorithm can be applied separately on that ONI file. Therefore consistent input data can be guaranteed and the performance of the SLAM approaches can be compared for one particular scenario. When algorithms are applied, duration and load are measured. Together with the results, the platform offers the functionality of exporting this data. Finally a modular comparison can be performed. Additionally, the position transformation are visualized in 2D and 3D.

To calculate the accuracy of the algorithms in different test environments, the paths are marked with tape and when passing one of those marks, the picture frame number is logged. Later the calculated position by the algorithms and the real position can be compared. To enable similar conditions between the test environments, the test person carrying the Kinect tries to hold the Kinect in the same manner for all walked paths in all scenarios.

IV. Evaluation results

We chose three different test environments to evaluate the performance of the algorithms in different scenarios and situations. The first test environment was a 7 room apartment, the second environment was an office building with connected rooms and the third environment was a subway station in Munich.

A. Test environments

In the **apartment scenario**, the visual key points approach was evaluated first. By comparing the SURF with the optical flow/KLT approach, the SURF approach outperforms the KLT (compare Figure 4). The effect of changes in the maximum distance of inliers to the model of the RANSAC algorithm were examined next. Comparing SURF and KLT, both approaches show similar effects. The best results are

 ISBN: 978-1-61208-203-5

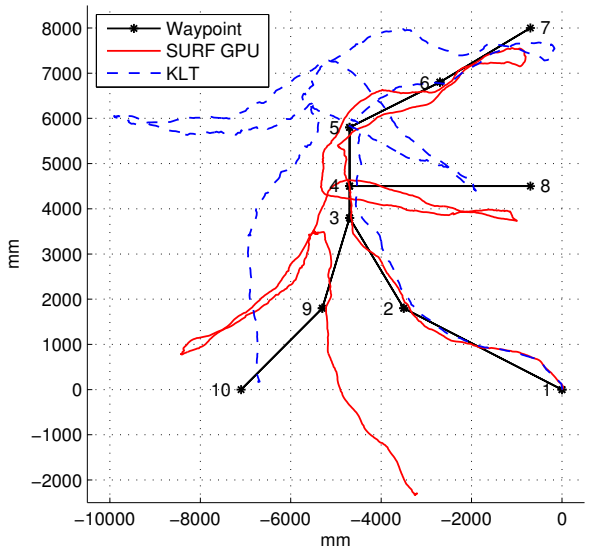


Figure 4. Resulting graphs of SURF and KLT

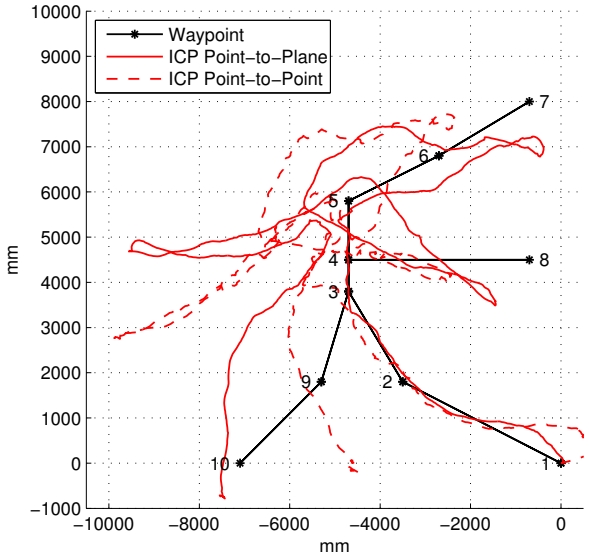


Figure 5. Point-to-Plane and Point-to-Point

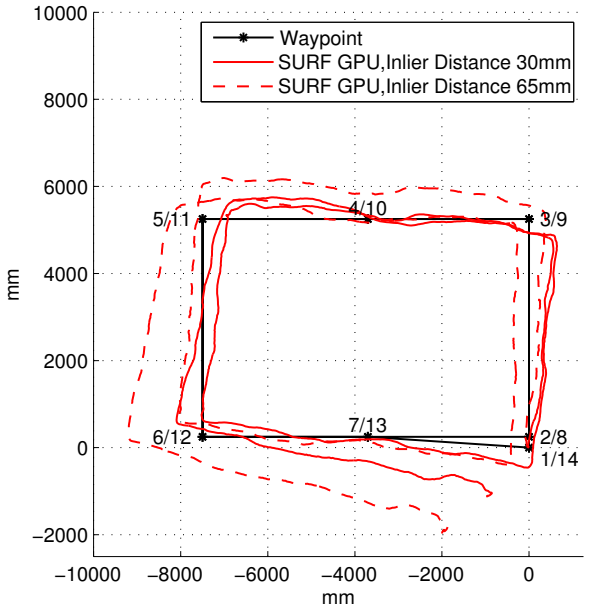


Figure 6. Two best Inlier distance values

achieved with a distance of 65 mm, higher or lower maximum values result in less accurate graphs. Applying the loop detection algorithm and TORO (where every 40th tracknode is compared to the new added) results in an enhancement of the graph. The enhancement for the KLT approach is higher than for the SURF approach. If more tracknodes are considered no significant enhancement could be measured.

Within the ICP method we compare the Point-to-Point method and the Point-to-Plane method. In Figure 5 the Point-to-Point method underlies a strong drift from the beginning on, whereas the Point-to-Plane method performs very well until the fifth waypoint. Afterwards, we examined the effect of different sizes of point clouds. Smaller variations of the size do not influence the Point-to-Plane method, whereas the Point-to-Plane method is sensitive to changes of the size. In comparison, the Point-to-Plane method is more robust and calculates good results with smaller point clouds.

In the **office scenario** the rooms were connected and the path walked outlines a closed rectangle. By varying the distance of the Inlier to the model for the RANSAC algorithm, similar results to the apartment scenario are calculated. The best two values for the maximum distance are depicted in Figure 6. Applying the loop detection algorithm and TORO results also in a enhancement of the graph. Interesting in this case is that the reduction of the track node distance from 40 to 20 in combination with the KLT and TORO, do not result in significant changes of graph accuracy. The evaluation of ICP algorithms (compare Figure 7) shows similar results to the apartment scenario.

The **subway station scenario** depicts a special scenario which differs in various aspects from the two previous scenarios. Subway stations consist of large areas and big halls. Since the range of the Kinect is limited, the test scenario was adjusted and the way the camera was positioned changed. To allow the Kinect to at least gather some depths information the Kinect was tilted towards the floor. Furthermore, bright illumination causes a lot of reflexions, which disturb the algorithms. After testing both visual methods and the ICP methods, the only approach which could calculate positions at all was the SURF approach. Both KLT and ICP method failed in the environment of the subway station.

B. Conclusions

Concluding the visual approaches, the SURF approach performed better than the KLT in all scenarios and test envi-

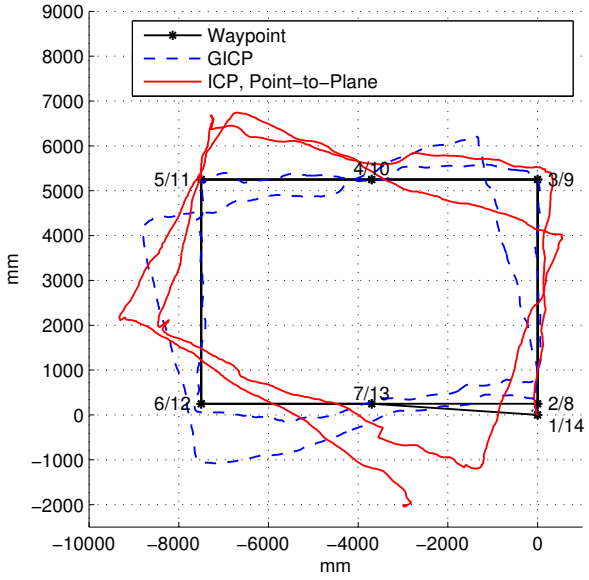


Figure 7. ICP approaches

ronments. In the test environment of the subway station, the KLT approach failed entirely because of variations of lighting, homogenous surfaces and missing depth information. TORO enhances both approaches in the apartment scenario, whereas in the office scenario SURF could be enhanced more with TORO than KLT. Varying the maximal distance between inliers and the model for the RANSAC algorithm enhanced both approaches. A standard configuration that performs equally well for all scenarios could not be found. For the SURF method, a maximal distance between 30 and 65 mm is feasible and for the KLT method between 30 and 50 mm.

Concluding the ICP approach, the Point-to-Plane Minimization method outperforms the Point-to-Plane method in the apartment scenario. An interesting aspect is the size of the point cloud. It was not the biggest point cloud that obtained more favorable results. In the Point-to-Plane alternative 3000 points achieved the best results.

Visual approaches could be further enhanced by inserting a refinement via ICP. The KLT approach reaches in each scenario the best performance in combination with the ICP, whereas for the SURF approach the data set and the test environment are crucial whether ICP can enhance the approach further more or not.

C. Overview of error rate

For the overview of the deviation in Figure 8, the best results from the apartment and the office test scenarios were accumulated and an average calculated. The scenario of the subway station was left out, since not all methods could provide feasible results.

An overview of accuracy, calculation time and robustness is given in Table I. The results of the subway scenario

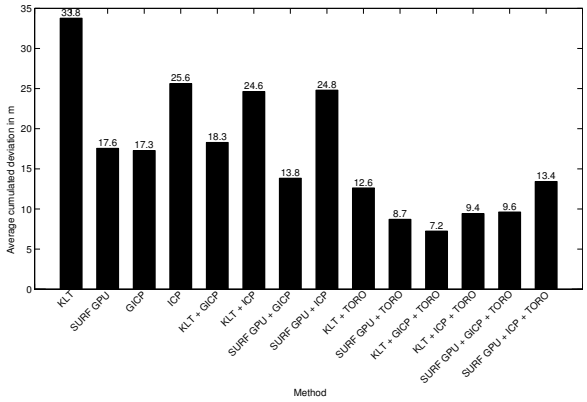


Figure 8. Average deviation (apartment and office tests)

were included in this overview. The subway scenario shows the weakness of the visual approaches. Since in subway stations the conditions are harsh (lighting changes extremely, homogenous surfaces and reflexions) key points could not always be found. The vast areas and big halls furthermore hamper SLAM methods using the Kinect.

V. Conclusion

In this work, we have shown that the Microsoft Kinect can be used for visual odometry and therefore is suitable for indoor positioning solutions in public buildings. For this purpose we tested the aptitude of state of the art techniques like SURF, RANSAC and ICP in combination with the Kinect in different scenarios. The results showed that every approach has some flaws, depending on the scenario.

For that reason, we developed a hybrid approach which makes use of visual methods as well as ICP. In order to do this, we use a customized RANSAC and then enhanced the results by additionally applying the ICP. For this purpose, we used a weight function customized for the Microsoft Kinect. All approaches were tested with an evaluation software which enabled us to test the approaches in real life environments and allowed us to record those environments for evaluations.

The results show that each the ICP and the hybrid approach usually outperform the pure visual methods inside of buildings. The scenario of the subway stations depicted a very harsh environment, where the sensors of the Kinect delivered weak data and only the SURF approach could estimate positions at all.

References

[1] C. Schindhelm, F. Gschwandtner, and M. Banholzer, “Usability of apple iphones for inertial navigation systems,” in *Proceedings of the 22nd Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Toronto, Canada, September 2011.

Method	∅ cummulated deviation (in m)	∅ calculating time (in ms)	robustness
KLT	33.7755	146.6980	-
SURF GPU	17.5606	106.5613	++
GICP	17.2786	273.3400	+
ICP	25.6255	200.6345	+
KLT + GICP	18.2869	342.7750	+
KLT + ICP	24.6325	310.9405	+
SURF GPU + GICP	13.8349	294.6580	++
SURF GPU + ICP	24.8067	282.1535	++
KLT + TORO	12.6179	635.1590	-
SURF GPU + TORO	8.7115	280.6925	++
KLT + GICP + TORO	7.2415	941.1915	+
KLT + ICP + TORO	9.4268	903.1700	+
SURF GPU + GICP + TORO	9.6246	538.9260	++
SURF GPU + ICP + TORO	13.4291	478.8190	++

Table I
OVERVIEW OF EVALUATION RESULTS

[2] H. Durrant-Whyte and T. Bailey, "Simultaneous localisation and mapping (slam): Part i the essential algorithms," *IEEE Robotics and Automation Magazine*, vol. 2, pp. 1–9, 2006.

[3] T. Bailey and H. Durrant-whyte, "Simultaneous localisation and mapping (slam): Part ii state of the art," *Computational Complexity*, vol. 13, no. 3, pp. 1–10, 2006.

[4] P. Robertson, M. Puyol, and M. Angermann, "Collaborative pedestrian mapping of buildings using inertial sensors and footslam," in *Proc. of the 24th International Technical Meeting of The Satellite Division of the Institue of Navigation (ION GNSS 2011)*, Portland, Oregon, September 2011, pp. 1366–.

[5] A. Garulli, A. Giannitrapani, A. Rossi, and A. Vicino, "Mobile robot slam for line-based environment representation," in *44th IEEE Conference on Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05.*, Seville, Spain, December 2005, pp. 2041–2046.

[6] X. S. Zhou and S. I. Roumeliotis, "Multi-robot slam with unknown initial correspondence: The robot rendezvous case," in *in Proceedings of IEEE International Conference on Intelligent Robots and Systems*, Beijing, China, October 2006, pp. 1785–1792.

[7] N. Engelhard, "Real-time 3d visual slam with a hand-held rgb-d camera," *Pattern Recognition*, vol. 2, no. c, 2011.

[8] The PrimeSense website. [Accessed Apr. 27, 2012]. [Online]. Available: http://www.primesense.com/

[9] The Willow Garage website. [Accessed Apr. 27, 2012]. [Online]. Available: http://www.willowgarage.com/

[10] The SideKick website. [Accessed Apr. 27, 2012]. [Online]. Available: http://www.sidekick.co.il/

[11] L. Gallo, A. P. Placitelli, and M. Ciampi, "Controller-free exploration of medical image data: Experiencing the kinect," in *Proceedings of the 24th IEEE International Symposium on Computer-Based Medical Systems, 27-30 June, 2011, Bristol, United Kingdom*. IEEE, 2011, pp. 1–6.

[12] M. Tolgyessy and P. Hubinsky, "The kinect sensor in robotics education," in *Proc. of 2nd International Conference on Robotics in Education (RiE 2011)*, R. Stelzer and K. Jafarmadar, Eds. Vienna, Austria: INNOC - Austrian Society for Innovative Computer Sciences, September 2011, pp. 143–146.

[13] W. K. Pratt, *Digital image processing*. New York, NY, USA: John Wiley & Sons, Inc., 1978.

[14] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the Sixth International Conference on Computer Vision*, ser. ICCV '98. Washington, DC, USA: IEEE Computer Society, January 1998, pp. 839–.

[15] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision–ECCV 2006*, pp. 404–417, 2006.

[16] D. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2. Kerkyra, Greece: IEEE, September 1999, pp. 1150–1157.

[17] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[18] G. Grisetti, C. Stachniss, and W. Burgard, "Non-linear constraint network optimization for efficient map learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, pp. 428–439, 2009.

[19] A. Segal, D. Haehnel, and S. Thrun, "Generalized-icp," in *Proc. of Robotics: Science and Systems*, 2009.

[20] G. Godin, M. Rioux, and R. Baribeau, "Three-dimensional registration using range and intensity information," in *Proceedings of SPIE*, vol. 2350, Boston, Massachusetts, November 1994, p. 279.

[21] Y. Chen and G. Medioni, "Object modeling by registration of multiple range images," in *Robotics and Automation, 1991. Proc., 1991 IEEE International Conference on*. Sacramento, California, USA: IEEE, April 1991, pp. 2724–2729.

 ISBN: 978-1-61208-203-5

Identifying Sources of Interference in RSSI Traces of a Single IEEE 802.15.4 Channel

Sven Zacharias, Thomas Newe, Sinead O'Keeffe, Elfed Lewis
Department of Electronic and Computer Engineering
University of Limerick
Limerick, Ireland
{Sven.Zacharias, Thomas.Newe, Sinead.OKeeffe, Elfed.Lewis}@ul.ie

***Abstract*—This paper presents the possibility of using RSSI readings to monitor a single IEEE 802.15.4 channel in the 2.4 GHz ISM band. An overview of the main sources of interference - namely Wireless Local Area Networks (WLANs), Bluetooth devices and microwave ovens - is given. Finally, an algorithm to classify one second of RSSI readings into one of these device classes is presented. The algorithm classifies 762 of 790 samples (96.46 %) correctly, having its worst precision with 97.41 % for the Bluetooth device class and its worst recall/sensitivity with 84.21 % for the microwave oven class. This algorithm gives an overview of interfering wireless devices without the need of changing the channel and thus allowing a continuous message reception.**

Keywords-IEEE 802.15.4; Radio Signal Strength Indicator (RSSI); 2.4 GHz ISM band; interference; coexistence; Wireless Sensor Network (WSN)

I. INTRODUCTION

Wireless Sensor Networks (WSNs) are small, embedded, in-expensive, low-power networks that are going to be widely deployed in the near future. They can be used in many applications in homes, offices and all sorts of urban environments. Today's most suitable wireless transfer technologies for WSNs are based on the IEEE 802.15.4 standard [1], since it provides a simple, low-power stack for the Physical and Medium Access Control (MAC) Layer. The IEEE 802.15.4 (2003) standard can physically operate in the three free Industrial, Scientific and Medical (ISM) frequency bands offering 27 channels: one at 868 MHz, ten in the 915 MHz band and 16 in the 2.4 GHz band. The only frequency band available worldwide is 2.4 GHz, which is the most used ISM band, utilized by many technologies and therefore the band is crowded [2]. Since wireless sensor nodes are power-constrained, energy saving by means of avoiding retransmissions or unnecessary on-times of the radio is an important task. Finding sources of interference allows avoiding collisions and therefore retransmissions can be reduced. This helps to have more reliable and energy efficient WSNs.

In the following section, a discussion of related work is given. Then the properties of Radio Signal Strength Indicator (RSSI) values are presented. Afterwards, the common sources of interference in WSNs are described, namely: Wireless Local Area Networks (WLANs), Bluetooth devices (BT) and microwave ovens (MWOs). For each device class, a short summary is given and then meaningful features for the detection are highlighted. Based on that, an algorithm is developed to identify the just mentioned device classes by RSSI readings of a single WSN channel. Subsequently, an evaluation and discussion of the algorithm is given. The paper ends with conclusions showing the potential fields of application for this work.

II. RELATED WORK

The coexistence of IEEE 802.15.4 with other IEEE standards has already been partly considered in the standard itself (Annex E). To avoid packet loss, the ZigBee standard recommends spectrum scanning with the help of RSSI readings for network channel management [3]. The scheme is only based on noise floor measurements on different channels and changes to a less used channel. There is no classification of sources of interference.

Boano et al. are using RSSI readings to improve the channel simulation [4] and to recreate interference [5]. Especially [5] gives a good overview of the possibilities of RSSI readings and the sources of interference (as in this work, WLAN, BT and MWO are researched). Emulations of the different sources of interference are presented, but no classification is used.

Rayanchu, Patro and Banerjee use an off-the-shelf WLAN interface card to measure the spectrum in the 2.4 GHz band and to identify devices [6]. Since IEEE 802.11 wireless network interface cards have different technical properties compared to IEEE 802.15.4 radios, their identification method differs from the one presented here. Their presented solution performs full spectrum scans and their classification of devices is based on a decision tree created with the help of machine learning.

Chowdhury and Akyildiz propose spectrum sensing with the help of a sensor node and an offline interference source classification approach. They scan the full spectrum and identify WLANs and MWOs by matching the observed spectral pattern with a stored reference shape. Their approach scans the full spectrum, thus the sensor node cannot receive while performing the scan. Their number of researched devices for WLANs and MWOs is rather small. They further suggest a scheme to choose the channel, packet scheduling times and sleep-awake cycles [7].

The algorithm presented here only needs the readings of a single channel and thus, the measuring sensor node is

 ISBN: 978-1-61208-203-5

connected to the network all the time. Also, the number of researched devices is high for an approach using sensor nodes.

III. RSSI READINGS

The IEEE 802.15.4 standard defines that an "Energy Detection" (ED) value must be measured for the "network layer as part of a channel selection algorithm. It is an estimate of the received signal power within the bandwidth of an IEEE 802.15.4 channel. No attempt is made to identify or decode signals on the channel. The ED time shall be equal to 8 symbol periods." This ED value is also widely known as the RSSI value. Since no identifying or decoding takes place, the RSSI can be used either to detect noise on a channel, or to indicate the quality of an incoming packet when measured while receiving.

Many applications and protocols for WSNs use the RSSI values to detect traffic or interference on the channel and to estimate transmission distances. Thus, RSSI is an enormously useful metric when used as a link quality estimator [8] or as part of a link quality estimator [9], and therefore, for routing. In addition, localization [10], channel management [7] and other systems rely heavily on RSSI readings.

In this work, RSSI readings from the Tmote Sky [11] sensor node are used. The data sheet of the built-in CC2420 radio chip [12] states a dynamic range from -100 to 0 dBm with an accuracy of ±6 dB and a linearity of ±3 dB. The RSSI is read over an 8 symbol period, which is 128 µs long in average. The quality of these RSSI readings was researched in [13] and the effects of the antenna pattern are shown in [14].

IV. SOURCES OF INTERFERENCE

The main sources of interference for WSNs in the 2.4 GHz band in urban environments and their effects on WSN deployments are reviewed in literature [15, 16]. In the literature and from the authors' experience, the main sources of interference are given as:

A. Wireless Local Area Networks

The term WLAN or Wi-Fi is commonly used to describe a collection of different technologies based on the IEEE 802.11 standard and its amendments [17]. In the following, the 802.11b, g and n standard [18] are of interest, since these operate in the 2.4 GHz band. Dependent on national restrictions there are up to a maximum of 14 (11 in North America) channels available. The IEEE 802.11b and g channels are 22 MHz wide and their channel center frequencies are only 5 MHz away from each other, thus they overlap each other. Channel 14 is an exception being 12 MHz away from its predecessor (see Figure 5). IEEE 802.11n works basically on the same channels but supports 40 MHz wide bundled channels and multiple-input multiple-output (MIMO), which is based on multiple antennas. Although spread spectrum modulated signals are used, a single 2 MHz wide WSN channel within the 22 MHz wide WLAN channel shows a clear peak in the RSSI readings of the Tmote Sky sensor node on WLAN sending activity. Hence the transmitting time and temporal length can be roughly detected. Since this work concentrates on single channel measurements the spectral properties cannot be used for identification.

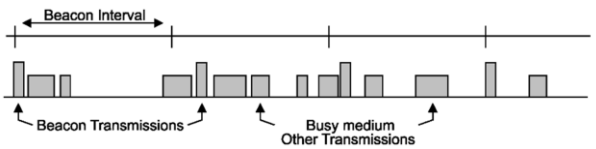


Figure 1. Beacon transmissions on a busy network [17].

The data rates of the previous mentioned standards are 1, 2, 11, 54 and 150 Mbit/s. Although there are different data rates, the standard specifies beacons, send by the Access Point (AP), which are different to normal traffic.

1) Beacon Frames: Every AP periodically sends a beacon frame to announce its network and to maintain connection to all clients in range. To allow all network interface cards to see the network, this beacon is send with the lowest data rate (1 or 2 Mbit/s) for highest compatibility. The smallest theoretical beacon has a body of around 30 bytes and 28 bytes of management frame. It has a measureable transfer time of roughly 0.5 or 0.25 ms. Most beacon frames are over 100 bytes in length and therefore, they are clearly traceable. The default behavior is to send ten beacons per second. The authors observed that all scanned WLANs (six in an office and 16 in a domestic environment) used a beaconing frequency of 10 Hz. This frequency is assumed for the remainder of this work. The beacons are good indicators of the presence of a WLAN on the channel and can be clearly seen in the RSSI readings as shown in Figure 6 (a). When the channel is heavily used the beacons become harder to identify as the standard does not provide reserved timeslot for beacons (see Figure 1 and Figure 6 (b)). This means that the AP has to access the communication medium by using the CSMA/CA algorithm as all participants do, resulting in the possibility of delayed beacons. As the AP is further away from the measuring sensor node, the beacons get increasingly lost in the data traffic.

2) Non-Beacon Frames: All other traffic in the WLAN can be, depending on the network possibilities, transferred at a higher speed and is therefore harder to identify as WLAN traffic. Some small packets can even be too fast to be measured using 11 kHz RSSI readings. This traffic has no dominant pattern, due to the various amounts of different protocols and applications.

B. Bluetooth Devices

BT [19] (IEEE 802.15.1) is designed to be a low-cost, medium-power, robust, short-range communication platform for Wireless Personal Area Networks (WPANs). It also operates in the 2.4 GHz band using 79 different 1 MHz wide channels (see Figure 5). It supports different sending classes with different sending powers. There are different versions of BT available, supporting different data rates up to 3

 ISBN: 978-1-61208-203-5

Mbit/s for Version 2.0 + Enhanced Data Rate (EDR) onwards. Since BT uses Adaptive Frequency Hopping (AFH) it is the least interfering technology presented here. BT changes the channel 1,600 times a second. This results in a time of 0.625 ms between the hops, called a slot, which is still traceable with a sampling rate of 11 kHz. A BT signal is characterized by its short spikes, due to the channel hops. The transmissions are organized by a Time Division Multiple Access (TDMA) scheme. BT supports two types of physical links: Synchronous Connection-Oriented (SCO) links and Asynchronous Connection-Less (ACL) links. SCO links are normally used for voice transfer and are strictly based on single slot packets. ACL links are packet based and can use one, three or five slots (see Figure 2). The traffic load and therefore the channel usage depend very much on the used application profile and wireless environment. The traffic can be low (regular traffic as for a wireless input device) to high (burst traffic as for file transfer (FTP)) or evenly spread transfer of audio as used for wireless headsets (see Figure 6 (d) and Figure 6 (e)). The actual transfer spikes of BT in the RSSI readings are the most reliable method for identification. The discovery and connection phase has not been investigated in this work.

C. Microwave Ovens

MWOs are a widely used household appliance working in the 2.4 GHz band with high power to warm food by dielectric heating. The common center frequency of MWOs is around 2.45 GHz with a spread width of at least 5 MHz and the average output power is around 800 W (the precise specification of a model can normally be found at the type plate at the back of the MWO). Through shielding most of the output power is kept in the cooking chamber of the device, but some waves are emitted to the environment. Measurements of the spectrum and the timing patterns of different MWOs can be found in [20].

MWOs consist of a single magnetron tube that emits high frequency waves. Since the magnetron works always with full power the user-set power level is achieved by controlling on and off periods. This results in off times between some heating phases (see Figure 3). These heating phases (shown in Figure 6 (c)) consist of wave emitting periods that are typically based on the frequency of the power supply (50 Hz in Europe or 60 Hz in North America). The periodical channel blocking differs very much to the signals used for digital, wireless communication and can be easily identified. For the rest of this paper, it is assumed that the MWO is

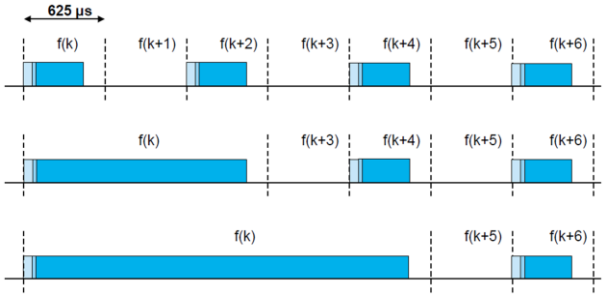


Figure 2. Single- and multi-slot packets used by Bluetooth [19].

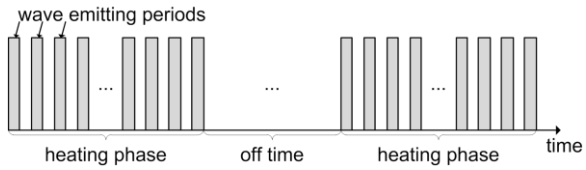


Figure 3. Simplified illustration of the wave emissions of a microwave oven operating with user setting “medium power”.

measured in a heating phase, because in the off times no waves are emitted.

D. Other Wireless Sensors Networks

Other WSNs operating on the same channel also have the potential to jam communications. The identification of other WSNs by RSSI readings would be possible (see Figure 6 (f)), but is not needed, since a single channel RSSI scanner can still receive messages. Even if the other WSN uses a different MAC protocol the message will still be received, but it might not be interpretable. Since the protocols used for WSNs are very variable, a time based classification based on RSSI readings would be quite complex to cover all possible patterns.

E. Other Devices

There are more devices active in the 2.4 GHz band, for example: Digital Enhanced Cordless Telecommunications (DECT) phones, wireless input devices not based on BT, or wireless video cameras, but they are beyond the scope of this paper.

V. IDENTIFYING DEVICES IN THE TIME DOMAIN

A. Experimental Setup

To develop a decision algorithm to identify the class of an interfering device, a data base of RSSI readings was created. All samples have been collected with a single Tmote Sky sensor node running ContikiOS 2.5 [21]. For measuring the Frossi Scanner [3, 22] has been used, recording RSSI readings with an average sampling rate of 11,321 Hz. With the help of MATLAB [23] 790 samples, each one second long, have been cut. These samples consist of scans of different channels in two WLAN environments, two MWOs, four BT devices, and another Tmote Sky sensor node sending short messages. All samples have been checked manually by viewing a plot to make sure that the sample is feasible and classifiable. This data base forms the foundation for the later stated detection rates. Its detailed composition is shown in Table 1.

B. Data Analysis

The main part of the data analysis was done offline in MATLAB. Additionally WEKA [24] was used, but the suggested trees and rules have not been used with the present algorithm, since they leak domain knowledge and are purely based on statistics. Some thresholds have been incorporated in the algorithm presented here.

 ISBN: 978-1-61208-203-5

```
// 1 second of RSSI readings
IF max(Readings) < 15
THEN return(NOISE);
IF ( (max(FFT.power).index between(48Hz,52Hz))
     OR
     (max(FFT.power).index between(98Hz,102Hz))
     AND (Usage between(30%,70%) )
THEN return(MWO);
IF ( (max(UsageLength) < 625 µs) OR
     (RaisingFlanks/count(RaisingFlanks)
      >= 0.286)) AND ((Usage <= 10%) AND
     (FFT.Power[10HZ]/sum(FFT.Power) <= 0.035) )
THEN return(BT);
IF ( (Usage between(1%,30%)) AND
     (max(ClearanceLengths) <= 100ms) )
THEN return(WLAN);
return(UNKNOWN);
```

Listing 1. Pseudocode of classification algorithm.

TABLE I. COMPOSITION OF THE USED DATA BASE OF RSSI READINGS.

Label	Type of device	Samples
WLAN	22 WLANs (partly overlapping, office and domestic environment)	640
MWO	2 different models of microwave ovens (manufacturers: Matsu, Bush)	19
BT	Laptop (Dell Wireless 370 Bluetooth Mini-card), Mobile Phone (Motorola Razr v3i), Headset (Samsung WEP-470), Wireless Mouse (Apple Magic Mouse)	121
WSN	Tmote Sky	10

C. Algorithm

The algorithm takes one second of RSSI readings as input and classifies it as WLAN, MWO, BT or unknown device. There is also the chance of an early return in the case where there is no signal present. In the following the algorithm is briefly described, an overview of the algorithm is given in Listing 1. The steps are worked through sequentially. If a classification matches, the result is returned and the algorithm ends.

1) Input: 1 s (~11,300 samples) of RSSI readings with values in the range of [0…100]. The dBm values can be computed as the RSSI values minus 100.

2) Noise: If no reading has a value greater or equal to 15, there is no classifiable signal present. In the following all values under 15 mean a free channel, while higher values are considered as usage of the channel. The default Clear Channel Assessment (CCA) threshold of the radio is 23. But the threshold of 15 allows the algorithm to work with weaker signals and is still far enough away from the noise floor.

3) MWO: The algorithm states that the signal is generated by a MWO if the following conditions are fulfilled: The maximum period power of the signal, found by a discrete Fourier transform, is between 48 and 52 or between 98 and 102 (based on European 50 Hz mean frequency). And the channel is used between 30 % and 70 % of the time.

4) BT: The algorithm states that the signal is generated by a BT device if the following conditions are fulfilled: The channel is never used longer than a single BT slot or the distance between rising flanks is mainly the [1…5] times of a slot time. And the channel is used less than 10 % of the time and the 10 Hz period power found by a discrete Fourier transform divided by the maximum power of all periods is less or equal to 0.035.

5) WLAN: The algorithm states that the signal is generated by a WLAN if the following conditions are fulfilled: The usage of the channel is between 1 % and 30 % and the maximum time of a clear channel is less than 100 ms (100 ms are the standard delay between to beacons).

6) UNKNOWN: If none of the previous conditions are fulfilled, the source of the signal is unknown.

D. Discussion of the Classification Results

The algorithm described performs well on the previously mentioned data base with 28 wrongly classified data sets out of 790 in total (3.54 %). The detailed confusion matrix is given in Table 2. Samples of WSNs were used to check the behavior of the classifier for unknown signals and to proof the exclusiveness of the classes, thus the precision value for WSNs is not meaningful. There is no class for WSNs since there is no need to detect them with RSSI readings (as explained in Section IV-D).

Since the signal is either binarized (channel used or clear) or normalized, as the FFT results, the distance to the interference source should be unimportant. The algorithm can be easily implemented on a personal computer. With an input of just one second of RSSI readings it is performed fast and can adapt quickly to changes in the wireless neighborhood. Unfortunately, at the moment, it is too complex to run on a sensor node. The memory of the node cannot handle the data.

The presence of multiple sources of interference is a challenge for the detection algorithm and the present algorithm only returns a single class. First trials showed that depending on the sources of interference different cases occur.

Since MWOs do not monitor or react to traffic on the medium, they overlay the signals of WLANs and BT devices and the algorithm will most likely not identify other sources of interference, due to the dominance of the MWO. The interference range of a MWO is quite limited, thus further away from the MWO the MWO signal will decline quickly and the other signals will become dominant and will be detected by the algorithm.

WLAN and BT are much more complex in there coexistence, since both of them react to the usage of the

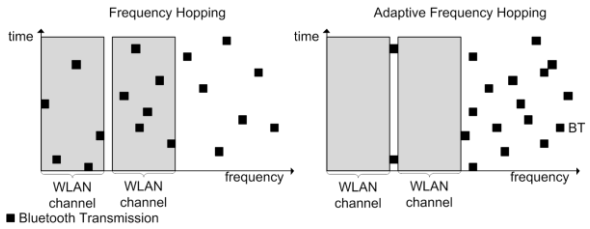


Figure 4. Simplified functional principle of Adaptive Frequency Hopping (AFH) compared to Frequency Hopping (FH).

medium. WLAN is quite widely spread, thus it can stand narrow band interference like BT. BT uses AFH and changes to un- or less used channels when many collisions occur on a channel. The principle of AFH is shown in Figure 4. Additionally, adaptive power control and Channel Quality Driven Data Rate (CQDDR) are used by BT to reduce interference. In the real world there are still some BT transmissions on the channels used by WLAN. But since there are many factors (distance to the sources of interference, data traffic and protocols used, and the just name interference avoiding technologies of BT) the interplay of BT and WLAN is not fully covered by the presented algorithm. In case of signals of BT and WLAN it will mostly classify the signal as WLAN, since WLAN is the dominant source of interference. According to [15], WLANs lead to much more lost packets than BT devices and so the algorithm returns the most relevant source of interference. Nevertheless, the detection of multiple sources of interference is a possible future enhancement for the presented algorithm.

New classes of devices as mentioned in Section IV-E, like wireless DECT phones and other proprietary devices operating in the 2.4 GHz band could be added. The algorithm needs further testing with more devices. Also the quality of the RSSI readings across different nodes of the same and different models could be compared. According to [13] the RSSI readings across different sensor nodes are comparable and hence the usage of other nodes as measurement devices is feasible.

As a short sanity check for the results, the authors went to another location and measured with a different Tmote Sky on channel 12 that was used by WLANs. All samples measured were classified correctly as WLANs. More consolidation of the results will be done in near future.

VI. CONCLUSIONS

This paper reviews the possibility of using RSSI readings to monitor the wireless channel. The main sources of wireless interference are introduced, and finally an algorithm to classify one second of RSSI readings into a device class is presented.

The results presented here can be used in many applications. The features of the signals highlighted here, can help to better simulate interference for improved channel models. The algorithm could run on the base station of a WSN enabling the base station to perform a funded centralized channel management. Channel sensing is also an important step for Cognitive Radios [25]. With the knowledge of the channel number the identification results could be improved further. An additional full spectrum scan could considerably improve the classification, but the ability to received messages without interruption on the channel would be lost. To the best of the authors' knowledge this is the first algorithm using a time series of RSSI readings of only a single channel to classify the wireless neighbors of a wireless sensor node.

ACKNOWLEDGMENT

The authors wish to thank the following for their financial support: The Embark Initiative and Intel, who fund this research through the Irish Research Council for Science, Engineering and Technology (IRCSET) postgraduate Research Scholarship Scheme.

REFERENCES

[1] *IEEE Standard 802.15.4™-2003*, IEEE Computer Society Std., 2003.

[2] G. Zhou, J. A. Stankovic, and S. H. Son, "Crowded spectrum in Wireless Sensor Networks," in *Proceedings of 3rd Workshop on Embedded Networked Sensors (EmNets)*, 2006.

[3] ZigBee Specification, ZigBee Alliance Std., January 2008.

[4] C. Boano, K. Römer, F. Österlind, and T. Voigt, "Demo abstract: Realistic simulation of radio interference in COOJA," in *Proceedings of the European Conference on Wireless Sensor Networks (EWSN)*, 2011.

[5] C. Boano, T. Voigt, C. Noda, K. Römer, and M. Zuniga, "Jamlab: Augmenting sensornet testbeds with realistic and controlled interference generation," in *10th International Conference on Information Processing in Sensor Networks (IPSN), 2011*, April 2011, pp. 175 –186.

[6] S. Rayanchu, A. Patro, and S. Banerjee, "Airshark: Detecting non-WiFi RF devices using commodity WiFi hardware," in *Internet Measurement Conference (IMC)*, Berlin, Germany, November 2011.

[7] K. Chowdhury and I. Akyildiz, "Interferer classification, channel selection and transmission adaptation for Wireless Sensor Networks," in *IEEE International Conference on Communications, 2009. ICC '09*, June 2009, pp. 1 –5.

[8] K. Srinivasan and P. Levis, "RSSI is under appreciated," in *In Proceedings of the Third Workshop on Embedded Networked Sensors (EmNets)*, 2006.

[9] N. Baccour, A. Koubaa, L. Mottola, M. Zuniga, C. Boano, and M. Alves, "Radio link quality estimation in Wireless Sensor Networks: A survey," *ACM Transaction on Sensor Networks*, 2011.

[10] A. Boukerche, H. A. B. F. Oliveira, E. F. Nakamura, and A. A. F. Loureiro, "Localization systems for Wireless Sensor Networks," in *Algorithms and Protocols for Wireless Sensor Networks*, A. Boukerche, Ed. John Wiley & Sons, Inc., Hoboken, New Jersey, 2009, pp. 307 – 340.

[11] Moteiv Corporation, *Tmote Sky - Ultra low power IEEE 802.15.4 compliant wireless sensor module*, 2006.

[12] *CC2420 2.4 GHz IEEE 802.15.4 / ZigBee-ready RF Transceiver*, Chipcon, June 2004, datasheet.

[13] Y. Chen and A. Terzis, "On the mechanisms and effects of calibrating RSSI measurements for 802.15.4 radios," in *7th European Conference on Wireless Sensor Networks (EWSN)*, 2010, pp. 256–271.

[14] M. Holland, R. Aures, and W. Heinzelman, "Experimental investigation of radio performance in Wireless Sensor Networks," in *2nd IEEE Workshop on Wireless Mesh Networks (WiMesh)*, IEEE, 2006, pp. 140–150.

[15] A. Sikora and V. Groza, "Coexistence of IEEE 802.15.4 with other systems in the 2.4 GHz-ISM-band," in *Proceedings of the IEEE Instrumentation and Measurement Technology Conference (IMTC)*, vol. 3, May 2005, pp. 1786 –1791.

[16] Jennic - Technology for a changing world, "Co-existence of IEEE 802.15.4 at 2.4 GHz," Application Note Revision 1.0, February 2008.

[17] *IEEE Standard 802.11™-2007*, IEEE Computer Society Std., 2007.

 ISBN: 978-1-61208-203-5

[18] *IEEE Standard 802.11n™-2009*, IEEE Computer Society Std., 2009.

[19] *Bluetooth Specification Version 2.1 + EDR*, Bluetooth SIG, Inc. Std., July 2007.

[20] P. Gawthrop, F. Sanders, K. Nebbia, and J. Sell, "Radio spectrum measurements of individual microwave ovens volume 1," NTIA Report 94-303-1, Tech. Rep., March 1994.

[21] A. Dunkels, B. Gronvall, and T. Voigt, "Contiki - a lightweight and flexible operating system for tiny networked sensors," in *29th Annual IEEE International Conference on Local Computer Networks*, IEEE, 2004, pp. 455–462.

[22] C. A. Boano, J. Eriksson, F. Österlind, and A. Dunkels "Contiki projects: Frossi scanner," http://contikiprojects.svn.sourceforge.net/viewvc/contikiprojects/sics.se/frossi-scanner/, [April 2012].

[23] MATLAB, *version 7.9.0.529 (R2009b)*. Natick, Massachusetts: The MathWorks Inc., 2009.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[25] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 1, pp. 116–130, 2009.

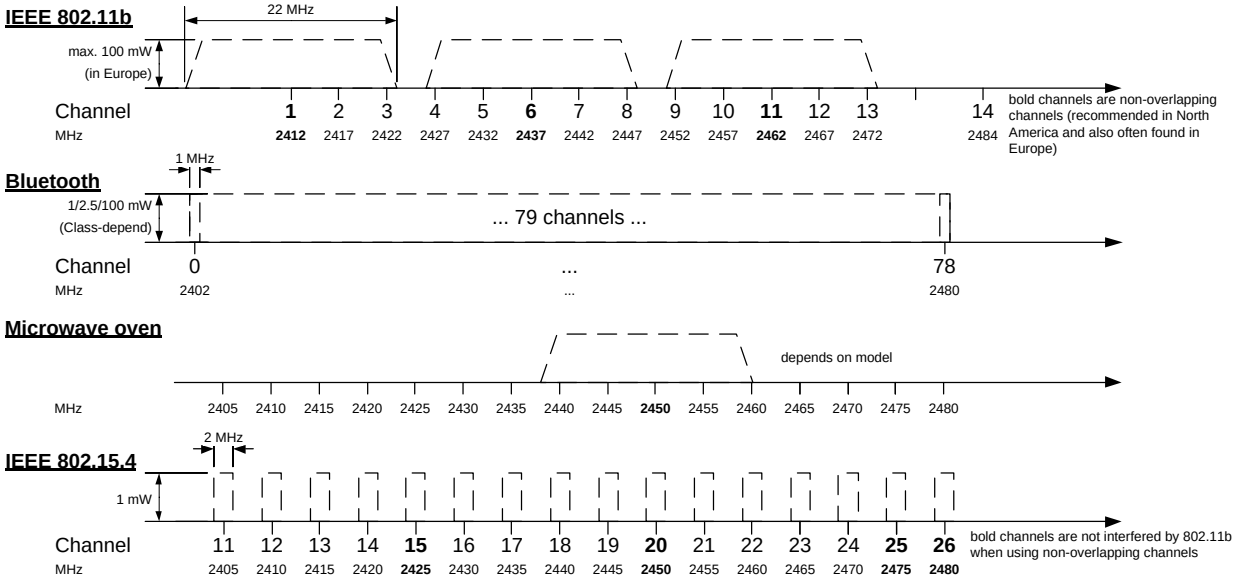


Figure 5. Overview of the usage of the 2.4 GHz spectrum by different standards/devices.
Do not scale spectral mask or output power from this drawing.

TABLE II. CONFUSION MATRIX OF IDENTIFIED CLASSES.

		Predicted class					
		WLAN	***BT***	***MWO***	***UNKNOWN***	***Precision***	***Recall/Sensitivity***
Actual class	***WLAN***	623	3	0	14	99.05 %	97.34 %
	BT	3	113	0	5	97.41 %	93.39 %
	MWO	3	0	16	0	100.00 %	84.21 %
	WSN	0	0	0	10	34.45 %	100.00 %

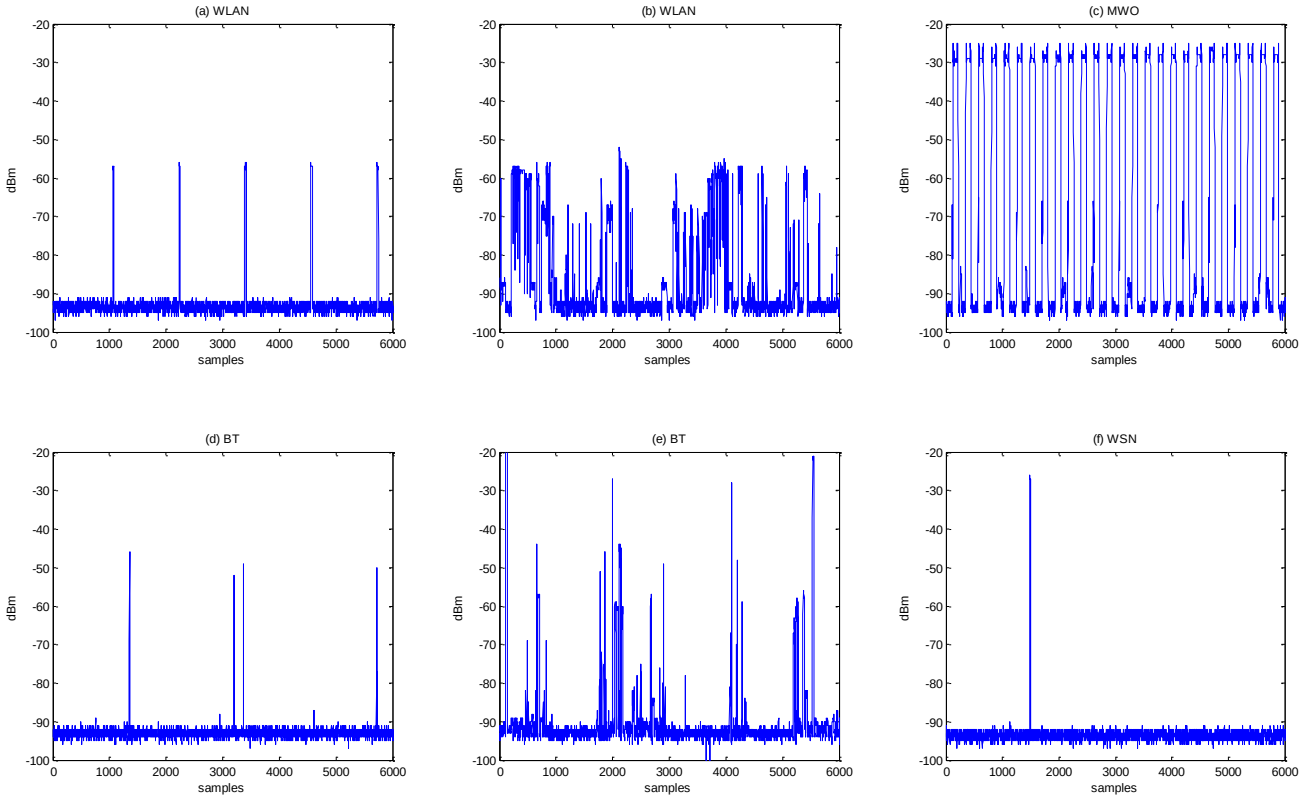


Figure 6. Overview of typical RSSI time series (0.5 s) of different devices. (a) WLAN beacons. (b) WLAN data traffic. (c) MWO heating phase. (d) Low traffic BT using only single-slot packets. (e) High traffic BT using multi-slot packets. (f) WSN sending a 22-byte-long message.

 ISBN: 978-1-61208-203-5

Dynamic Distributed Resource Allocation in Relay Assisted OFDMA Networks

Javad Hajipour[1], Amr Mohamed[2], and Victor C. M. Leung[1]
[1]*ECE Department, University of British Columbia, Vancouver, Canada.* {*hajipour, vleung*}*@ece.ubc.ca*
[2]*CSE Department, Qatar University, Doha, Qatar.* {*amrm*}*@qu.edu.qa*

***Abstract*—Relay assisted OFDMA networks are promising solutions for provision of high-data-rate services in wide coverage areas. However, the deployment of relays makes the resource allocation a more challenging and complex task. In this paper we study dynamic allocation of power and subchannels in an OFDMA downlink system with regenerative relays which have the capability of buffering the users' data to transmit in a suitable time. We model the network as a multicell scenario with small serving areas and provide a novel framework for resource allocation, in which each of the relays and Base Station (BS) allocate resources based on the queue and channel state information of their own users. We propose a dynamic distributed resource allocation algorithm for this purpose, where BS and relays decide about the allocation of the power and subchannels by passing messages among themselves and based on the local queue and channel state information. Simulation results show significant improvement in terms of system throughput and users' queue stability.**

***Keywords*-OFDMA; regenerative buffering relays; dual decomposition; distributed resource allocation.**

I. INTRODUCTION

Orthogonal frequency division multiple access (OFDMA) is a promising solution for multiple access in high speed wireless networks such as IEEE 802.16 Worldwide Interoperability for Microwave Access (WiMAX), and Long Term Evolution (LTE). Based on this technique, it is possible to provide high spectaral efficiency, multiuser diversity, robustness against multipath fading and flexibility in radio resource allocation. However to make it possible for all the users in a large area to get access to the network, wide coverage is another important objective for next generation of mobile networks. For this purpose, wireless relays have gained significant attention in both industry and research bodies, due to their cost effective and fast deployments. Relays can improve the transmission link between Base Station (BS) and users which are far from BS or have blockage between BS and themselves.

Resource allocation and scheduling are important issues in wireless networks due to increasing demand of users for data traffic and the scarcity of radio resources [1]. It becomes even more challenging and crucial in relay assisted OFDMA networks[2]. Recently there has been remarkable work done in this field [3][4][5][6]. In [3], authors studied the capacity of relay assisted OFDMA networks for both amplify-and-forward (AF) and decode-and-forward (DF) schemes. Adaptive scheduling algorithms have been studied in [4] and, based on a Time Division Duplexing (TDD) frame structure, Greedy Polling (GP) and Partial Proportional Fair (PPF) algorithms have been proposed. In [5] authors assumed that the frequency band is partitioned between users connected directly to BS and users connected through relays. They studied the cross layer scheduling for the relayed users in an AF relay network, as an optimization problem with the objective of maximizing the received goodput and proposed a distributed algorithm for it. Most of the works in this area use two common assumptions. First one is that users have infinitely backlogged buffers in BS, meaning that they always have data to transmit; However in realistic scenarios, this assumption is not true and users have random and bursty traffic arrival of packets which feed users' buffers in BS. Therefore the channel aware scheduling without considering the availability of data, would lead into inefficient use of resources. The other common assumption is that relays are "prompt" and the relaying is performed in two consecutive transmission epochs [4]. In other words, time slots are assumed to be divided into two subslots where in the first one, BS transmits to the relays and in the second one, they forward the received data to their users. However having relays with the capability of buffering data and forwarding them in a later time, can provide more flexibility for Radio Resource Management (RRM) as it is possible to keep a user's data in the queue and forward it in a suitable time slot, i.e., when the user's channel gets better or the user gets higher priority. Such a system has been considered in [6], and based on that authors have studied the joint routing and subchannel allocation in a relay assisted OFDMA cellular network. They have considered concurrent transmission for relays where a relay can receive data on some subchannels and at the same time transmit on some others. Assuming equal power allocation and equal number of frequency subchannels being used by either of BS and any of the relays, a centralized algorithm has been proposed. However optimal power allocation is another important factor for efficient utilization of the system resources and providing Quality of Service (QoS) for users in terms of Bit Error Rate (BER) and queue stability [7][8].

In this paper, we consider a relay assissted OFDMA network with buffering capability in relays and availibility of all of the subchannels to all of the BS and Relays. We formulate joint channel and power allocation as an optimization problem and introducing some concepts, we

 ISBN: 978-1-61208-203-5

show its similarity to a multicell OFDMA scenario with smaller cells. Moreover, to make the problem tractable, we transform it into a convex optimization problem and using dual decomposition, we propose an iterative Dynamic Distributed Resource Allocation (DDRA) algorithm, where BS and each relay solve their own problem based on their users' Queue and Channel State Information (QCSI) and some global variables exchanged among them. DDRA provides a novel framework for exploiting the system's power and subchannel resources in an effeicient and adaptive way over time, with lower overhead of the CSI feedback and lower computational complexity at the BS compared to optimal centralized scheduling which requires global CSI at the BS.

The rest of the paper is organized as follows. In Section II, we outline the model for the relay assisted OFDMA system. In Section III, we formulate the resource allocation algorithm design as an optimization problem and solve it by dual decomposition, where distributed closed-form solutions for power and subcarrier allocation are derived. Simulation results for the distributed algorithm are studied in Section IV with conclusion finally presented in Section V.

II. System Model

We consider a single cell time slotted OFDMA system in downlink (DL) with K users and M relays. Users are uniformly distributed in the cell, K_1 of them being served directly by BS while others receive data through one of the relays. As it is shown in Figure 1, we assume that each user has been assigned to either BS or any of the relays based on a criteria such as average Signal to Nosie Ratio (SNR), distance from the BS and relays, etc.

Relays' locations are fixed and can have different distances from BS, based on the topology of the service area. BS and relays are equipped with buffers, where BS has one for each user but relays have one for each of only the users connected to them. Users' packets arrive at the BS buffer according to their traffic model and are queued until transmission to the directly connected users or to relays serving other users. Relays do not need to transmit the received packets immediately in the next time slot and it is possible to keep them in the buffers and serve them based on the scheduling policy. This gives flexibility to the scheduler to utilize the resources more opportunistically by postponing the transmission until the user gets higher priority or better channel. We use Q_k^B, $k = 1, \ldots, K$ to denote the queue size of user k in BS, and $Q_k^{R(k)}$, $k = K_1 + 1, \ldots, K$ to denote the queue size of user k in its serving relay, $R(k)$.

We assume that transmission bandwidth is divided into N subchannels where each subchannel can be used exclusively by BS or relays in one of the groups of the links, i.e., BS-to-users, BS-to-relays and relays-to-users. Any relay has the ability to transmit on some subchannels and at the same time receive data from BS on other ones. The channels in all the links are assumed time variant and frequency selective, but

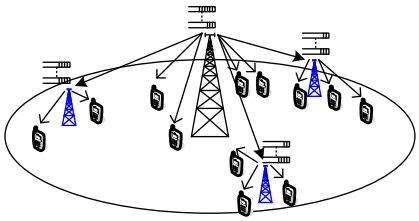


Figure 1. System model

constant during one time slot. We define the gain-to-noise ratio corresponding to the link between BS and user k as follows:

$$e_{kn}^B = \frac{|H_{kn}^B|^2 G_k^B}{\sigma_n^2}, \quad (1)$$

where H_{kn}^B is the small scale fading coefficient between BS and user k in subchanel n, G_k^B is the path loss attenuation between BS and user k and σ_n^2 is the variance of Gaussian noise. $e_{kn}^{R(k)}$ and $e_n^{BR(k)}$ can be defined in a similar way for the links between $R(k)$ and user k and the links between BS and $R(k)$. Assuming that M-ary QAM modulation is used for transmission, the achievable transmission rate can be computed as follows[9]:

$$r_{kn}^B = x_{kn}^B \log_2 \left(1 + \frac{p_{kn}^B e_{kn}^B}{\Gamma_k}\right), k = 1, \ldots, K_1, \quad (2)$$

where, without loss of generality, the bandwidth of a subchannel has been assumed equal to 1. r_{kn}^B is the achievable transmission rate between BS and user k on subchannel n. x_{kn}^B denotes subchannel allocation indicator which would be one if subchannel n is used by BS to transmit data to user k, $k = 1...K_1$, and zero otherwise. p_{kn}^B is the power allocated by BS to user k on subchannel n. Γ_k is the SNR gap due to the limited number of coding and modulation schemes and is related to bit error rate of user k (BER_k), through equation $\Gamma_k = -\frac{\ln(5BER_k)}{1.5}$. In a similar way we can define $x_{kn}^{R(k)}$,$p_{kn}^{R(k)}$ and $r_{kn}^{R(k)}$ for the links of relays-to-users and $x_{kn}^{BR(k)}$, $p_{kn}^{BR(k)}$, and $r_{kn}^{BR(k)}$ for the links of BS-to-relays.

III. Cross Layer Scheduling and Resource Allocation

In this section, we formulate the cross layer scheduling and resource allocation and then using some definitions and modifications, we propose a new perspective with simplified convex optimization problem.

A. Problem Formulation

In each time slot, the resource allocation policy aims at efficient use of the system resources, i.e., power and subchannles, while considering the QoS for the users, in terms of BER and queue stability. For this purpose, a weight is considered for each of the users on their links and the objective is to maximize the weighted throughput over the

 ISBN: 978-1-61208-203-5

links. The cross layer scheduling and resource allocation can be formulated as the following optimization problem:

$$P: \max_{\boldsymbol{p},\boldsymbol{x}} \sum_{k=1}^{K_1}\sum_{n=1}^{N} w_k^B r_{kn}^B + \sum_{k=K_1+1}^{K}\sum_{n=1}^{N} w_k^{BR(k)} r_{kn}^{BR(k)} + \sum_{k=K_1+1}^{K}\sum_{n=1}^{N} w_k^{R(k)} r_{kn}^{R(k)}, \quad (3a)$$

$$\text{s.t. } C1: \sum_{n=1}^{N}\Big(\sum_{k=1}^{K_1} p_{kn}^B + \sum_{k=K_1+1}^{K}(p_{kn}^{BR(k)} + p_{kn}^{R(k)})\Big) \le P_t, \quad (3b)$$

$$C2: \sum_{k=1}^{K_1} x_{kn}^B + \sum_{k=K_1+1}^{K}(x_{kn}^{BR(k)} + x_{kn}^{R(k)}) \le 1, \forall n, \quad (3c)$$

$$C3: x_{kn}^B, x_{kn}^{BR(k)}, x_{kn}^{R(k)} \in \{0,1\}, \forall k, n, \quad (3d)$$

$$C4: p_{kn}^B, p_{kn}^{BR(k)}, p_{kn}^{R(k)} \ge 0, \forall k, n \quad (3e)$$

where w_k^B, $w_k^{BR(k)}$, and $w_k^{R(k)}$ are the weights of the users over the links of BS-to-users, BS-to-relays and relays-to-users. Constraint $C1$ is the total power constraint for the BS and the relays. The problem (3a) is a complex combinatorial optimization problem, which needs an exhaustive search to find the optimal solution. In order to make the problem tractable, we relax the subchannel assignment variables $x_{kn}^B, x_{kn}^{BR(k)}, x_{kn}^{R(k)}$ to be real value between zero and one, instead of a Boolean, i.e., $0 \le x_{kn}^B, x_{kn}^{BR(k)}, x_{kn}^{R(k)} \le 1$, which is known as time or tone sharing. Furthermore, we consider the buffers of users $k = K_1, \ldots, K$ in their relays, as virtual users that are directly connected to BS. In other words, we interpret the links between BS and relays as the direct links between BS and some virtual users. As shown in Figure 2, this perspective helps us to divide the serving area (single cell) into smaller areas (multi cells) served by $M+1$ nodes, where node 1 is BS with K users and has the complicated RRM capability and act as a central controller while nodes $m, m = 2, \ldots, M+1$, are the relays with their own users, totally $K - K_1$ users, and acts as antennas distributed in the serving area and connected wirelessly to the controller. We denote the set of users of node m with $\mathcal{U}_m$; in particular $\mathcal{U}_1 = 1...K$. Each node has the buffers of its own users and transmits data independently; however in the beginning of each slot they all communicate with a central controller in node 1 to decide about their shares of power and subchannels in order not to make interference to other nodes.

We use the following notations for each user:

$$e_{kn}^m = \begin{cases} e_{kn}^B, & m=1, k=1,\ldots,K_1 \\ e_n^{BR(k)}, & m=1, k=K_1+1,\ldots,K \\ e_{kn}^{R(k)}, & m=2,\ldots,M+1, k=K_1+1,\ldots,K, k\in\mathcal{U}_m \end{cases}$$

w_k^m, x_{kn}^m and p_{kn}^m can be defined in a similar way. We define $\mathcal{D} = \{(\boldsymbol{p},\boldsymbol{x}) | 0 \le p_{kn}^m \le P_t, x_{kn}^m \in [0,1]\}$ as the domain of the problem. Due to tone sharing, SNR will be equal to $\frac{p_{kn}^m e_{kn}^m}{x_{kn}^m}$;

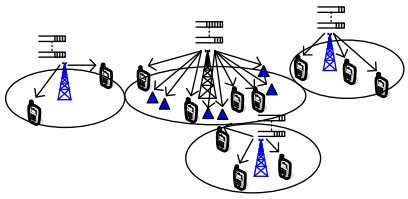


Figure 2. Similarity of the model to multicell network

this SNR is because of viewing p_{kn}^m as the energy per time slot that node m uses for user k on subchannel n [10]. As a result the rates will be computed by $r_{kn}^m = x_{kn}^m \log_2(1 + \frac{p_{kn}^m e_{kn}^m}{x_{kn}^m \Gamma_k})$. Assuming that the system is stabilizable, similar to [11], it can be proved that queue stability can be provided by defining the weights of users as follows:

$$w_k^m = \begin{cases} Q_k^B, & m=1, k=1,\ldots,K \\ Q_k^{R(k)}, & m=2,\ldots,M+1, k=K_1+1,\ldots,K, k\in\mathcal{U}_m \end{cases} \quad (4)$$

Considering these weights for queue stability provision, makes it possible for BS and relays to utilize only local QCSI for resource allocation algorithm provided in the subsequent subections. Using the framework mentioned above, resource allocation problem can be represented as follows:

$$\max_{\boldsymbol{p},\boldsymbol{x}\in\mathcal{D}} \sum_{m=1}^{M+1}\sum_{k\in\mathcal{U}_m}\sum_{n=1}^{N} w_k^m x_{kn}^m \log_2(1 + \frac{p_{kn}^m e_{kn}^m}{x_{kn}^m \Gamma_k}), \quad (5a)$$

$$\text{s.t. } C1: \sum_{n=1}^{N}\sum_{m=1}^{M+1}\sum_{k\in\mathcal{U}_m} p_{kn}^m \le P_t, \quad (5b)$$

$$C2: \sum_{m=1}^{M+1}\sum_{k\in\mathcal{U}_m} x_{kn}^m \le 1, \forall n \quad (5c)$$

It is worth to note that the ordinary OFDMA networks can be considered as a special case of this formulation where M=0; in that case the virtual users will become the real users directly connected to BS.

Problem 5 is convex and the strong duality holds [10] (This can be verified by defining $\tilde{p}_{kn}^m = \frac{p_{kn}^m}{x_{kn}^m}$ and substituting in the objective and constraints). Therefore, using dual decomposition, an iterative algorithm can be designed to solve the problem.

B. Dual Problem Formulation

In this subsection, we formulate the dual problem for the resource allocation optimization problem. For this, we first obtain the Lagrangian function of primal problem. After

 ISBN: 978-1-61208-203-5

rearranging the terms, the Lagrangian can be written as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{p},\boldsymbol{x},\mu,\boldsymbol{\delta}) &= \sum_{m=1}^{M+1}\sum_{k\in\mathcal{U}_m}\sum_{n=1}^{N} w_k^m x_{kn}^m \log_2(1+\frac{p_{kn}^m e_{kn}^m}{x_{kn}^m\Gamma_k}) \\ &- \sum_{m=1}^{M+1}\sum_{k\in\mathcal{U}_m}\sum_{n=1}^{N}\mu p_{kn}^m \\ &- \sum_{m=1}^{M+1}\sum_{k\in\mathcal{U}_m}\sum_{n=1}^{N}\delta_n x_{kn}^m \\ &+ \mu P_t + \sum_{n=1}^{N}\delta_n \end{aligned} \quad (6)$$

where μ is the Lagrangian multiplier associated with total power constraint and $\boldsymbol{\delta}$ is the Lagrangian multiplier vector for the subchannel allocation constraints. The dual problem is given by:

$$\min_{\boldsymbol{\mu},\boldsymbol{\delta}\geq 0}\ \max_{\boldsymbol{p},\boldsymbol{x}\in\mathcal{D}} \mathcal{L}(\boldsymbol{p},\boldsymbol{x},\mu,\boldsymbol{\delta}) \quad (7)$$

Similar to the method in [10], the dual problem can be solved by a centralized iterative algorithm in BS. In this case, since the BS has the information of the previous transmissions, it would have QSI of all the relays, but it will need to ask for the CSI for the links between relays and their users on all the subchannels which will lead to an overhead in the order of O($(K-K_1)N$). Alternatively, as in [5], using dual decomposition and concept of pricing, we propose an iterative distributed algorithm where in each iteration, BS and relays, solve their own problem based on the global variables and their local QCSI.

In the following subsection, we solve the dual problem in (7) by decomposing it into two parts: the first part is the local subproblem to be solved by each of the serving nodes, BS and relays, and the second part is the main dual problem to be solved by BS.

C. Dynamic Distributed Resource Allocation - DDRA

By dual decomposition, the dual problem is decomposed into a main global problem and $M+1$ local problems which can be solved iteratively. In each iteration, using the dual variables, which are global for all the nodes, BS and relays solve their local subproblem based on their QCSI. Then relays report their results to the BS and BS updates the dual variables and broadcasts them to relays. In this way, dual variables act as prices that BS adjusts to control the demands. The local subproblem in each node is given by:

$$\max_{\boldsymbol{p},\boldsymbol{x}\in\mathcal{D}} \mathcal{L}_m(\boldsymbol{p},\boldsymbol{x},\mu,\boldsymbol{\delta}), \qquad \text{with}$$

$$\begin{aligned}\mathcal{L}_m(\boldsymbol{p},\boldsymbol{x},\mu,\boldsymbol{\delta}) &= \sum_{k\in\mathcal{U}_m}\sum_{n=1}^{N} w_k^m x_{kn}^m \log_2(1+\frac{p_{kn}^m e_{kn}^m}{x_{kn}^m\Gamma_k}) \\ &- \sum_{k\in\mathcal{U}_m}\sum_{n=1}^{N}\mu p_{kn}^m \\ &- \sum_{k\in\mathcal{U}_m}\sum_{n=1}^{N}\delta_n x_{kn}^m \end{aligned} \quad (8)$$

where the lagrange multipliers μ and $\boldsymbol{\delta}$ are provided by the BS. Using the Karush-Kuhn-Tucker conditions we have:

$$\frac{\partial\mathcal{L}_m}{\partial p_{kn}^m} = \frac{w_k^m x_{kn}^m e_{kn}^m}{x_{kn}^m\Gamma_k + p_{kn}^m e_{kn}^m} - \mu = 0 \quad (9)$$

As a result, power allocation for subchannel n is obtained by:

$$\begin{aligned} p_{kn}^{m\,*}(\boldsymbol{x},\mu,\boldsymbol{\delta}) &= x_{kn}^m \tilde{p}_{kn}^m(\mu), \qquad \text{with} \\ \tilde{p}_{kn}^m(\mu) &= \min\Big(P_t, \Big(\frac{w_k^m}{\mu} + \frac{\ln(5BER_k)}{1.5\ e_{kn}^m}\Big)^+\Big) \end{aligned} \quad (10)$$

where $(a)^+$=max$(a,0)$. After substituting $p_{kn}^{m\,*}$ into (8), we have:

$$\mathcal{L}_m\ (\boldsymbol{x},\mu,\boldsymbol{\delta}) = \sum_{k\in\mathcal{U}_m}\sum_{n=1}^{N} x_{kn}^m V_{kn}^m \quad (11)$$

$$\text{with}\quad V_{kn}^m = w_k^m \log_2(1+\frac{\tilde{p}_{kn}^m e_{kn}^m}{\Gamma_k}) - \Big(\mu\tilde{p}_{kn}^m + \delta_n\Big)$$

Defining $V_n^{m*} = \max_{k\in\mathcal{U}_m}\{V_{kn}^m\}$, (11) will be maximized if subchannel assignment variables are computed as follows:

$$x_{kn}^{m\,*}(\mu,\boldsymbol{\delta}) = \begin{cases} 1, & V_{kn}^m = (V_n^{m*})^+, \\ 0, & V_{kn}^m < (V_n^{m*})^+, \end{cases} \quad (12)$$

In some time slots, more than one users might have $V_{kn}^m = (V_n^{m*})^+$. This happens mostly for the virtual users of BS that represent the links belonging to the same group, i.e., between BS and a particular relay, as these links have the same channel condition over a subchannel. In such cases, subchannel is allocated to the user that has larger queue size or better channel condition. According to (4), (10) and (12), queue sizes of users, their channel conditions and required BER will affect their share of power and subchennals.

D. Solution of Main Dual Problem at BS

Based on the information of the powers and channel allocation variables reported by relays and using subgradient method, BS will update the dual variables through the following iterations:

$$\begin{aligned}\mu(t+1) &= \Big[\mu(t) - \xi_1(t)\Big(P_t - \sum_{m=1}^{M+1}\sum_{k\in\mathcal{U}_m}\sum_{n=1}^{N} p_{kn}^m\Big)\Big]^+, \\ \delta_n(t+1) &= \Big[\delta_n(t) - \xi_2(t)\Big(1 - \sum_{m=1}^{M+1}\sum_{k\in\mathcal{U}_m}\sum_{n=1}^{N} x_{kn}^m\Big)\Big]^+, \forall n \end{aligned} \quad (13)$$

 ISBN: 978-1-61208-203-5

In this algorithm, the overhead of messages reported by relays will be in the order of $O(NM)$ multiplied by the number of iterations, which would be considerably lower than that of centralized algorithm in the networks with high number of users. Number of iterations can be optimized to reach fast convergence, by choosing suitable step sizes and initial values [5].

IV. Numerical Results

To evaluate the system performance, we have considered a system with M=3 and N=20 and have conducted extensive Matlab simulations over 1000 time slots. Simulation parameters are shown in table I. For the links from BS or relay to users, Rayleigh channel model is used, while the links from BS to relays, are modeled with Rician channel with κ factor equal to 6 dB. Results are presented in terms of system throughput as well as average and maximum queue sizes in the system. For baseline, we have used the PPF method proposed in [4] in which power is equally allocated over subchannels and relays are prompt, i.e., they transmit in a time subslot immediately after the reception subslot. We have adjusted PPF for our scenario by considering the availability of data in the queues of users in BS; we call it Queue Aware PPF (QAPPF), as it computes the achievable rates of users based on their queue size and channel conditions. Figure 3 shows the average system throughput in each time slot, with two values of systme total power. It is observed that DDRA is able to utilize the wireless resources more efficiently, compared to QAPPF. The reason is that although both of them share the system power and subchannels for the users directly connected to BS and users connected through relays, in DDRA BS and relays allocate power and subchannels adaptively and also have flexibility over time to transmit to users and as a result they are able to get higher benefit from resources and from time diversity. Also it is observed that as the number of users increases, DDRA is able to utilize the multiuser diversity and get more gain. This is also displayed in Figure 4, by the Cumulative Distribution Function (CDF) of throughputs in each time slot, in the case of 10 users. The jumps in the diagram of DDRA are because of the fact that it utilizes resources efficiently and is able to empty the queues sometimes. Then when a new packet is arrived in a time slot in one of the queues, it is transmitted completely. It is clear frrom Figure 4 that DDRA is able to provide higher bit rates with higher probability.

Table I
Simulation Parameters

Parameter Name	Setting
Cell Radius	1000m
Min UE-BS distance	50m
BS Antenna Height	15m
Relay Antenna Height	5m
User Antenna Height	1.5m
Relay Distance from BS	2/3 cell radius
Pathloss Model	From [12]
Subchannel Bandwidth	15 kHz
Time Slot Duration	1ms
BER Requirement	1e-6
Traffic Model/Packet Size	Poisson/1Kb
Packet Interarrival Time	30ms

Figure 5 demonstrates the average queue size in the system over time, with 10 users. While DDRA keeps queues stable, QAPPF is not able to reach this goal and therefore queue sizes increase unboundedly. We also show the CDF of maximum queue sizes in a system with 10 users in Figure 6. QAPPF results in higher probability for large queue sizes in the system, which would cause higher probability of buffer overflow. On the other hand, DDRA is able to keep queue sizes in smaller ranges and guarantee system stability. This is due to the fact that according to (4) DDRA gives higher weights to the users with larger queue sizes and using (10) and (12), it is able to allocate resources adaptively based on queue size, channel condition and required BER of the users. As the results confirm, DDRA is a throughput optimal algorithm, meaning that it is able to keep the queue sizes bounded if it is feasible at all [11].

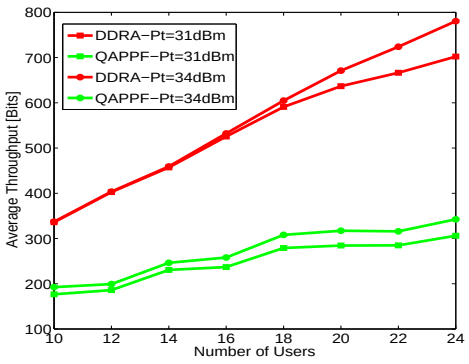


Figure 3. System Average Throughput

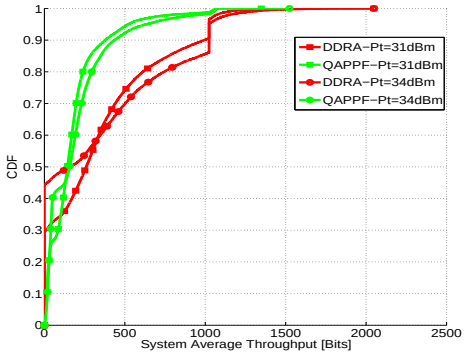


Figure 4. Distribution of System Average Throughput, K=10

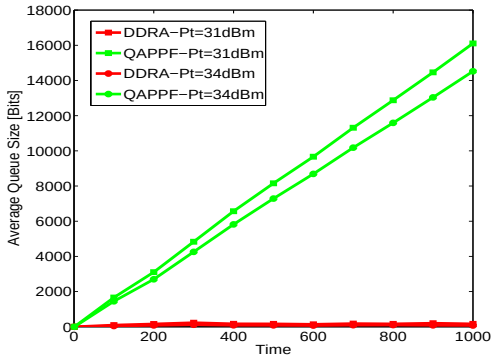


Figure 5. System Average Queue Size Over Time, K=10

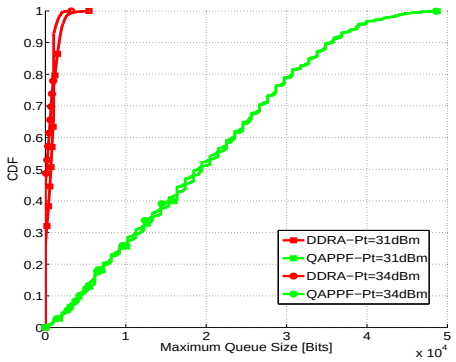


Figure 6. Distribution of System Maximum Queue Size, K=10

V. Conclusion and Future Work

In this paper we provided a novel framework for joint power and subchannel allocation in a relay assisted OFDMA network, with the assumption that relays are able to buffer data and transmit in a later time. Defining the links between BS and relays as virtual users, a new perspective was provided and similarity of the system to a multicell network was shown. We formulated the resource allocation problem as a convex optimization problem and using dual decomposition, we proposed an iterative Dynamic Distributed Resource Allocation (DDRA) algorithm, in which each of the BS and relays solve their own problem based on some global variables and their local information about queue and channel states of their users. The closed form equations derived for power and subchannel allocation, reveals the adaptive characteristic of our resource allocation algorithm based on queue size, channel condition and required BER of the users. The propsed perspective and algorithm, is highly scalable which is of great appeal for deployment and radio resource management of relay assisted OFDMA networks. Numerical results confirm the throughput optimality of DDRA and show significant improvement in the system performance in terms of average throughput and queue stability. As the future work, we will consider separate power constraints for BS and relays and will extend DDRA for a scenario with both delay-sensitive and delay-tolerant services.

Acknowledgement

This work is supported in part by the Qatar Telecom (Qtel) Grant no. QUEX-Qtel-09/10-10, the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Institute for Computing, Information and Cognitive Systems (ICICS) at UBC.

References

[1] M. Shariat and A. Quddus and S. Ghorashi, and R. Tafazolli, “Scheduling as an important cross-layer operation for emerging broadband wireless systems,” IEEE Communications Surveys and Tutorials, vol. 11, pp. 74-86, 2009.

[2] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, and D. Falconer, “Opportunities and challenges in OFDMA-based cellular relay networks: a radio resource management perspective,” IEEE Trans. Veh. Technol., vol. 59, pp. 2496-2510, Jun. 2010.

[3] G. Li and H. Liu, “On the capacity of broadband relay networks,” in 38th Asilomar Conference on Signals, Systems and Computers, vol. 2, pp. 1318-1322, Nov. 2004.

[4] L. Huang, M. Rong, L. Wang, Y. Xue, and E. Schulz, “Resouce scheduling for OFDMA/TDD based relay enhanced cellular networks,” in Proc. IEEE Wireless Commun. and Networking Conf., pp. 1544-1548, Mar. 2007.

[5] D. Ng and R. Schober, “Cross-layer scheduling for OFDMA amplify-and-forward relay networks,” IEEE Trans. Veh. Technol., vol. 59, pp. 1443-1458, Mar. 2010.

[6] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, and Y.-D. Kim, “Fairness-aware radio resource management in downlink OFDMA cellular relay networks,” IEEE Trans. Wireless Commun., vol. 9, pp. 1628-1639, May 2010.

[7] C. S. Park and K. B. Lee, “Transmit power allocation for BER performance improvement in multicarrier systems,” IEEE Trans. Commun., vol. 52, pp. 1658-1663, Oct. 2004.

[8] M. Neely, E. Modiano, and C. Rohrs, “Power allocation and routing in multibeam satellites with time-varying channels,” IEEE/ACM Trans. on Networking, vol. 11, pp. 138-152, Feb. 2003.

[9] X. Qiu and K. Chawla, “On the performance of adaptive modulation in cellular systems,” IEEE Trans. Commun., vol. 47, pp. 884-895, Jun. 1999.

[10] R. A. J. Huang, V. G. Subramanian and R. A. Berry, “Downlink scheduling and resource allocation for OFDM systems,” IEEE Trans. Wireless Commun., vol. 8, pp. 288-296, Jan. 2009.

[11] D. Park, “A throughput-optimal scheduling policy for wireless relay networks,” in IEEE Wireless Communications and Networking Conference (WCNC), pp. 1-5, Apr. 2010.

[12] “Spatial channel model for multiple input multiple output (MIMO) simulations,” 3GPP TR 25.996 V7.0.0 (2007-06), Tech. Rep.

 ISBN: 978-1-61208-203-5

A Protoyping Platform for Spectrum Sensing in China

Christian Kocks[1], Alexander Viessmann[2], Peter Jung[3]
Department of Communication Technologies
University of Duisburg-Essen
47057 Duisburg, Germany
[1]christian.kocks@kommunikationstechnik.org
[2]alexander.viessmann@kommunikationstechnik.org
[3]peter.jung@kommunikationstechnik.org

Lei Chen
Communication Technology Research Dept.
Huawei Tech. Co., Ltd.
hwchenlei@huawei.com

***Abstract*—Due to the increasing demand on wireless communications the idea of cognitive radio is of utmost interest. The TV white space may become the first commercial application of cognitive radio resulting from its advantageous propagation properties. It allows the usage of secondary communication systems at non-occupied frequency bands. Within this manuscript, a prototyping platform for cognitive radio applications is presented. Its underlying architecture is based on a combination of DSP and FPGA and relies on the software-defined radio paradigm. Spectrum sensing algorithms are introduced for the three predominant Chinese TV standards DTMB, CMMB and PAL-D/K. Finally, the algorithms' performance is shown in a comparison to simulation results. The focus of this manuscript is on a TV white space prototyping platform and the validation of spectrum sensing algorithms for the Chinese TV standards DTMB, CMMB and PAL-D/K.**

***Keywords*—CMMB; Cognitive Radio; DTMB; Prototyping Platform; PAL-D/K; TV White Space**

I. INTRODUCTION

In the recent decade, an increasing interest in the field of cognitive radio (CR) for wireless communication systems could be discovered. It is considered as a key technology for significantly alleviating the spectrum scarcity. One application for the CR technology is the TV white space (TVWS). It refers to non-occupied frequency bands in the TV spectrum, i.e., below 900 MHz, and is a desirable target for CR-based spectrum sharing due to its advantageous propagation properties compared to other frequency ranges on the one hand and due to its low utilization ratio on the other hand [1]. Hence, CR in TVWS will probably become the first commercial application that brings CR from concept to reality. In the United States, the FCC has already made an official request to allow unlicensed users reusing TV bands without causing interference to incumbent users [2]. In other countries, the corresponding regulatory authorities such as the CEPT in the European Union are developing regulations on the unlicensed usage in TVWS as well. Besides the regulatory authorities, standardization organizations such as IEEE 802.22 [3] have started the standardization for cognitive radio applications.

The spectrum sensing technology has been considered as a key element of CR and its application to TVWS has been widely studied. However, a variety of different TV standards exists, which may differ from country to country, especially for digital TV standards. While in North America ATSC (Advanced Television Systems Committee) is deployed, in Europe, South Asia and Africa, DVB-T/H (Digital Video Broadcasting - Terrestrial/Handheld) plays the predominant role. Further standards such as ISDB (Integrated Services Digital Broadcasting) developed in Japan or DMB (Digital Multimedia Broadcasting) developed in Korea are also used in various countries [4]. As a result, it is hardly feasible to design a universal sensing algorithm for all TV standards. This manuscript focuses on spectrum sensing for Chinese TV standards.

There are mainly three terrestrial and handheld TV standards in China: DTMB (Digital Terrestrial Multimedia Broadcast) [5] for terrestrial reception, CMMB (China Mobile Multimedia Broadcasting) [6] for handheld reception and PAL-D/K (Phase Alternating Line) [7] for analog TV. While other countries such as the USA have already stopped the provision of analog TV, the nationwide switchover from analog to digital TV will not occur until the year 2015. Therefore, the analog TV will still coexist with the digital TV for many years to come. As a result, the detection of both analog and digital signals is necessary for CR implementations.

The United States are the first and also most active country in exploiting the unlicensed usage of TVWS. The spectrum sensing technology for ATSC signals has been intensively studied. Several detection algorithms for ATSC and its analog predecessor NTSC (National Television System Committee) can be found in IEEE 802.22 standard [8]. In 2008, a sensing prototype test campaign was organized by FCC [9]. As an example, Motorola, Philips and I2R have submitted their prototype designs, which have been tested both in the laboratory as well as in the field. The results showed that the ATSC and NTSC signals can be detected correctly with a certain probability. As another widely used TV standard, DVB-T has also been intensively studied with respect to spectrum sensing. In [10], a robust sensing approach is discussed using a prototype sensor developed by Philips. Several detection algorithms for a Chinese standard, i.e., DTMB, have also

 ISBN: 978-1-61208-203-5

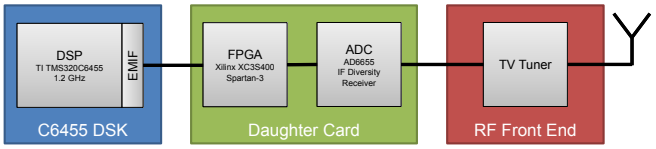


Fig. 1. Platform Overview for Spectrum Sensing Applications

been studied and published [11], [12]. The focus of this manuscript is on a prototyping platform developed by the authors and deployed for the implementation of spectrum sensing algorithms for the TVWS in China. The prototyping platform is based on the software-defined radio paradigm [13] allowing a reconfiguration of the platform by software. Besides the prototyping platform, spectrum sensing algorithms for DTMB, CMMB and PAL-D/K are illustrated including their measured performance in comparison with simulation results.

This manuscript is structured as follows. In Section II, the prototyping platform is presented, while, in Section III, the signal flow for the spectrum sensing operation is addressed. Section IV gives an overview of the Chinese TV standards DTMB, CMMB and PAL-D/K. The corresponding sensing algorithms, which are implemented on the prototyping platform, are presented in Section V. Section VI shows selected results in a comparison between the simulated algorithms' performance and the performance measured with the prototyping platform. Finally, a conclusion is given.

II. Spectrum Sensing Prototyping Platform

For the implementation of cognitive radios, an elaborated prototyping platform is essential. Already during the concept phase of this prototyping platform, modularity has been a crucial design constraint. Hence, the platform is designed in a way that certain parts can easily be replaced by more appropriate parts depending on the system to be implemented and its underlying requirements. The platform mainly consists of three printed circuit boards as illustrated in Figure 1. The base board is a DSP starter kit hosting a powerful DSP TMS320C6455 by Texas Instruments running at 1.2 GHz. This DSP is responsible for major parts of the signal processing algorithms on the one hand and for the overall platform scheduling on the other hand. Since higher-level programming languages such as C/C++ can be used for implementing signal processing algorithms, this DSP-based platform is ideally suited for rapid prototyping. Algorithms, which have been studied in a simulation environment before, can easily be implemented to run on the DSP. Furthermore, the advanced debugging capabilities of this DSP simplifies locating implementation errors.

Directly attached to the DSP is the mixed-signal daughter card. While in Figure 1 the core components required for the spectrum sensing operation are illustrated only, Figure 2 shows a more detailed block diagram of this daughter card. The daughter card can, besides the implementation presented here, also be used as a full communication transceiver. It consists

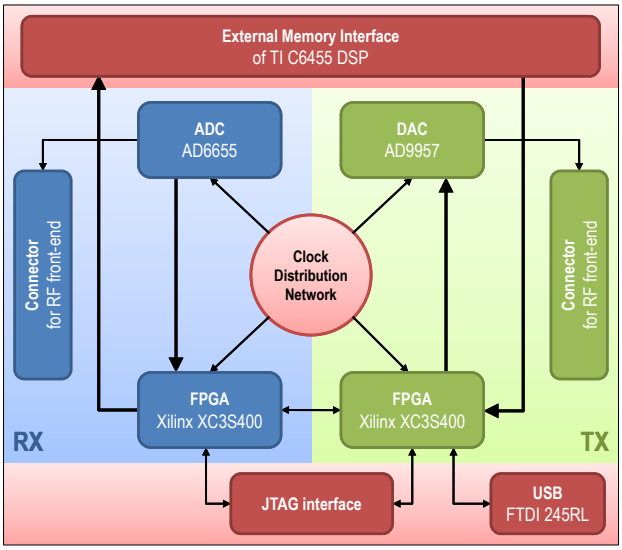


Fig. 2. Prototyping Platform: Block Diagram of the Mixed-Signal Daughter Card

of an Analog Devices AD6655 analog-to-digital converter (ADC) as well as of an Analog Devices AD9957 digital-to-analog converter (DAC). Both devices are supplied from an elaborate clock distribution network, which guarantees highly stable clocks for the overall platform. Since the focus of this manuscript is on the spectrum sensing implementation, in the following only the receiver branch of this daughter card is considered. The digitized signal coming from the ADC is directly given to a Xilinx Spartan-3 FPGA for performing further filtering and decimation operations. Furthermore, the FPGA is used for synchronizing the spectrum sensing events. The synchronization information originates from a Huawei LTE eNodeB. After some fundamental signal processing steps, the data is buffered within the FPGA and transferred to the DSP using the Texas Instrument EMIF (External Memory Interface). To reduce the overall load of the DSP, this transfer makes use of direct memory access (DMA).

The analog input signal of the ADC originates from an RF front-end directly attached to the mixed-signal daughter card. In case of the spectrum sensing prototyping platform, the RF front-end mainly consists of a commercially available TV tuner receiving the RF signal by an appropriate antenna and down-converting it to an intermediate frequency (IF) signal, which is then sampled by the ADC.

A photography of the spectrum sensing prototyping platform is given in Figure 3. It shows the three aforementioned modules with the RF front-end at the top and the DSK at the bottom. In between, the PCB of the daughter card is located. Additionally, a separate PCB is located on the right-hand side for debugging purposes and for interfacing with the synchronization entity.

 ISBN: 978-1-61208-203-5

III. Prototyping Platform Signal Flow

This section describes the signal flow for the spectrum sensing operation. The focus is on the digital baseband signal, which is buffered in the DSP. An overview of the signal flow gives Figure 4. Before the sensing operation starts, its parameters such as sensing interval, target false-alarm probability and TV standards to be sensed for are defined by an external spectrum management entity. Within the prototyping platform itself, a control unit is responsible for evaluating and distributing the parameters of interest. There are different operation modes depending on the a-priori knowledge about the underlying TV usage. In case the frequency band of interest may only be used by one TV standard, this information is communicated to the control unit so that only the corresponding detection algorithm is carried out. Otherwise, in case this frequency band may be used by all of the available TV standards, the control unit passes the captured data first to the DTMB detector followed by the CMMB detector and, finally, the PAL-D/K detector. The soft-decision outputs of all detectors are then processed by a combination metric to give an overall information about the presence of any of these signals. A graphical user interface (GUI) exists, which allows a simple configuration of the sensing parameters and an immediate demonstration of the sensing results.

IV. Chinese TV Standards

The intention of this section is to give a brief overview of the various Chinese TV standards. The focus is on the main aspects, which are relevant for feature-based signal detection. For a full description of the TV systems, please refer to [5], [6] and [7], respectively.

A. DTMB

DTMB, also referred to as DMB-T (Digital Multimedia Broadcast - Terrestrial), is a mandatory TV standard in China. DTMB can be used in either single-carrier or in multi-carrier mode. Three FEC (Forward Error Correction) code rates, five modulation orders and two interleaving depths are specified for DTMB [5]. A block diagram of a DTMB transmitter is shown in Figure 5.

DTMB defines three different header types with different lengths. The frame body itself has a fixed length of 500 µs. The frame structure of DTMB including the different header types is illustrated in Figure 6. The frames are hierarchically

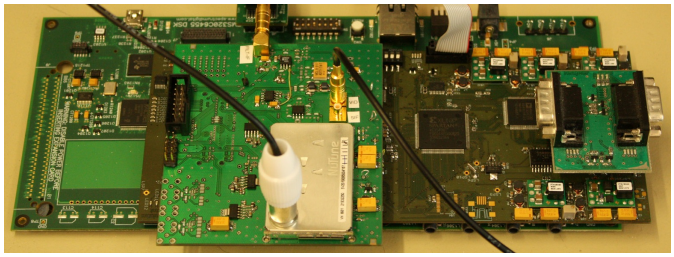


Fig. 3. Photography of the Prototyping Platform

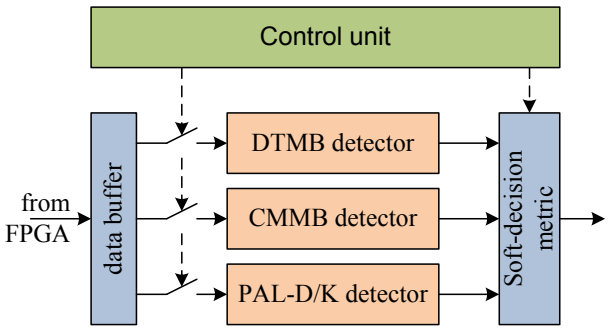


Fig. 4. DSP Signal Flow

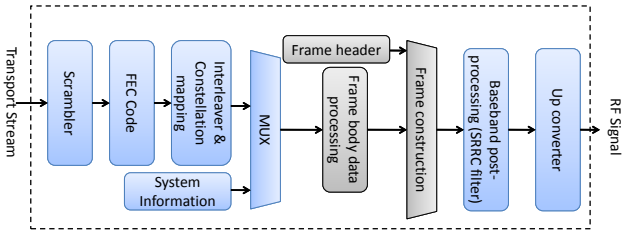


Fig. 5. DTMB Transmitter [5]

structured in a calendar day frame, a minute frame and a super frame. One superframe consists of either 225 frames with frame header mode 1 or of 216 frames with frame header mode 2 or of 200 frames with frame header mode 3.

The three frame headers are generated by different generator polynomials [5], which are:

$$G_1(x) = 1 + x + x^5 + x^6 + x^8 \quad (1)$$

for mode 1,

$$G_2(x) = 1 + x^3 + x^{10} \quad (2)$$

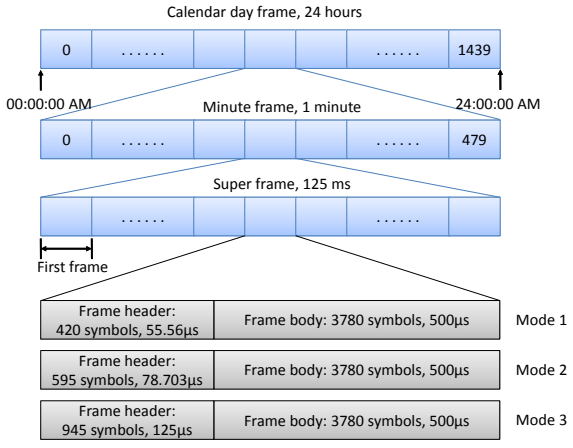


Fig. 6. DTMB Frame Structure [5]

 ISBN: 978-1-61208-203-5

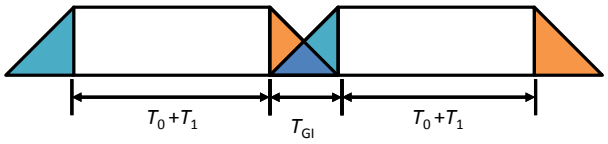


Fig. 7. CMMB Symbol Overlapping

for mode 2 and

$$G_3(x) = 1 + x^2 + x^7 + x^8 + x^9 \quad (3)$$

for mode 3. The generation of the sequence can be realized by a linear feedback shift register.

B. CMMB

CMMB is a system fully based on the well-known OFDM (Orthogonal Frequency Division Multiplexing). A combination of Reed-Solomon (RS) and Low-Density Parity-Check (LDPC) codes is used for FEC. Unlike many other OFDM systems such as DVB-T [14], the OFDM symbol of length T_0 in time-domain is not only extended by inserting a cyclic prefix (length T_1) but it is also extended by a pre-guard interval and a post-guard interval of length T_{GI} each. As illustrated in Figure 7, the post-guard interval of a certain OFDM symbol in CMMB overlaps with the pre-guard interval of the subsequent symbol [6].

In CMMB, one frame has a duration of 1 s and consists of 40 time slots. Each time slot contains one beacon and 53 OFDM symbols. The beacon contains a transmitter identification field and two synchronization signals. The OFDM symbols consist of data-bearing subcarriers as well as of pilot subcarriers. These pilot subcarriers are subdivided into continual pilots and scattered pilots [6].

C. PAL-D/K

A variety of different PAL-based standards exist, which mainly differ in the channel bandwidth or in the underlying modulation scheme. The PAL standard used in China is called PAL-D/K with 8 MHz channel bandwidth, 50 Hz field frequency and 625 lines per frame [7]. A PAL signal consists of separate video and audio parts. Within this manuscript, only the bandwidth occupied by the video part is subject to spectrum sensing. The video signal used in PAL is a CVBS (Color Video, Blanking and Sync) signal, which is an extension to the monochrome VBS (Video, Blanking and Sync) signal. A snapshot of a standard VBS signal is depicted in Figure 8. In addition to the video signal itself, the VBS signal has some additional components, which are required, e.g., for synchronization at the receiver. The black-level signal components after and before the video signal are referred to as front porch and back porch, respectively. The time values shown in Figure 8 are compliant to the PAL-D/K standard. The total duration of one line is $64\,\mu$s resulting in a line frequency of 15625 Hz [7].

V. Spectrum Sensing Algorithms

After the brief introduction to the various Chinese TV standards, this section describes the spectrum sensing algorithms. All algorithms have in common that they are based on autocorrelation of the digital baseband signal. In general, the autocorrelation function $\varphi_{ss}(t)$ of a complex signal $s(t)$ is defined as

$$\varphi_{ss}(\tau) = \int_{-\infty}^{\infty} s^*(t)s(t+\tau)dt \quad (4)$$

where $(\cdot)^*$ denotes the complex conjugation.

A. DTMB

In DTMB, the frame header appears periodically at the beginning of each frame, which can be exploited for the sensing operation. The presented autocorrelation-based sensing algorithm for DTMB can be divided into three stages:

- autocorrelation stage
- comb-correlation stage
- decision stage

A flow diagram of the algorithm is shown in Figure 9. For the autocorrelation the digitized baseband signal is multiplied with a delayed and complex conjugated version of the signal where the delay itself depends on the frame header mode. The running average filter cumulates a certain number of the multiplication output samples. Resulting from the periodical appearance of the frame header, the first stage's output is applied to a comb correlator, which is a correlation with a Dirac comb $g(t)$ with a distance Δt corresponding to the frame header period, i.e.,

$$g(t) = \sum_{k=-\infty}^{\infty} \delta(t - k \cdot \Delta t). \quad (5)$$

This stage allows collecting the energy of all frames within the sensing period. The squared magnitude of the cumulated comb-correlation output φ_{cc} is given to the decision stage. In this stage, the ratio λ of the maximum and the average of the

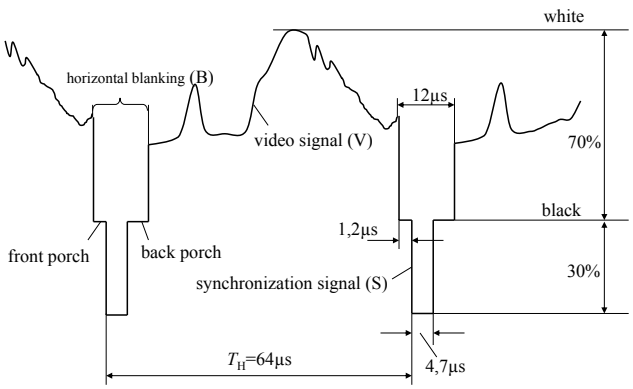


Fig. 8. VBS Signal

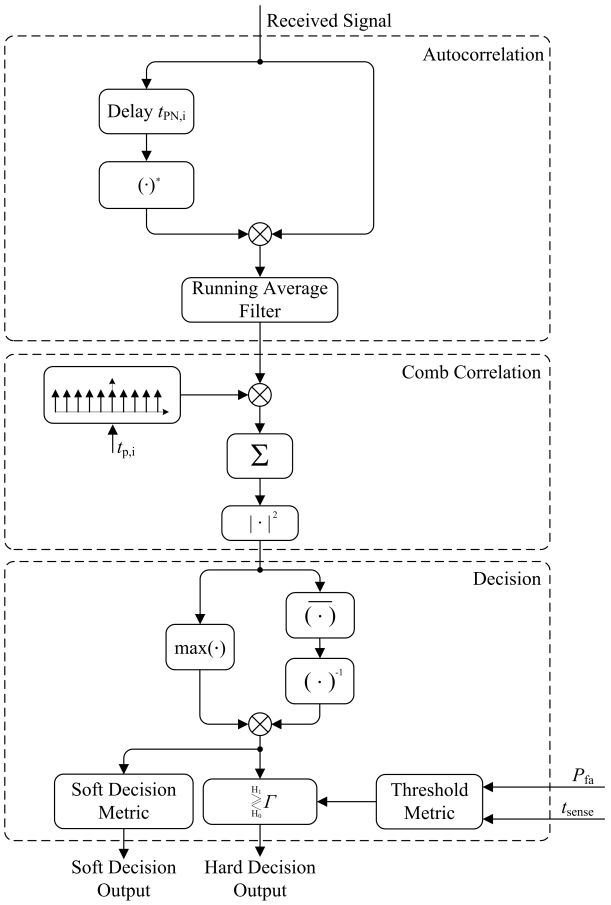


Fig. 9. Flow Diagram of the DTMB Sensing Algorithm

previous stage's output is calculated:

$$\lambda = \frac{\max\left(\varphi_{\mathrm{cc}}(t)\right)}{\overline{\varphi_{\mathrm{cc}}(t)}}. \tag{6}$$

By applying a soft-decision metric to λ, a measure for the probability of the presence of a DTMB signal is generated. Furthermore, comparing the ratio with a threshold Γ gives a hard-decision on the presence of a DTMB signal. This threshold is generated by using a threshold metric based on the available sensing interval t_{sense} and the desired false-alarm probability P_{fa}.

By using the ratio λ for making the decision about the presence of a DTMB signal, the presented algorithm is robust against dynamic range variations as well as varying signal-to-noise ratios and, thus, independent of the underlying AGC (Automatic Gain Control) implementation.

B. CMMB

The sensing algorithm for CMMB is very similar to the sensing algorithm for DTMB. As shown in Section IV, CMMB uses a cyclic repetition of certain parts of the OFDM symbol, denoted as cyclic prefix. Since this cyclic prefix equals the last part of the corresponding OFDM symbol, it is well suited for the sensing operation. The general data flow of the algorithm is identical to the DTMB algorithm depicted in Figure 9. However, the timings must be adapted according to the CMMB parameters.

C. PAL-D/K

The sensing algorithm for PAL-D/K relies on the periodicity of certain parts of the CVBS signal as depicted in Figure 8. The CVBS signal exhibits a periodic pattern of the synchronization pulses in every transmitted line of the resulting TV picture. In addition to the synchronization pulses itself with a length of $t_{\mathrm{hsync}} = 4.7\ \mu$s the front as well as the back porch with lengths of $t_{\mathrm{fp}} = 1.2\ \mu$s and $t_{\mathrm{bp}} = 6.1\ \mu$s, respectively, can be used for sensing purposes. The time between two consecutive synchronization pulses is $t_{\mathrm{H}} = 64\ \mu$s. A flow diagram of the PAL-D/K sensing algorithm is depicted in Figure 10 and consists of two stages:

- Autocorrelation stage
- Decision stage

While the delay in the autocorrelation corresponds to the periodicity of the CVBS signal, the length of the running average filter is set to t_{H} as well. This improves the sensing performance by exploiting similarities in the video signal for consecutive lines. In the decision stage, the average of the output of the autocorrelation stage is calculated. The residual parts of the decision stage are identical to the corresponding parts in the decision stages for DTMB and CMMB signals.

VI. Results

To show the overall performance of the previously introduced algorithms, a comparison between the simulation results and the results measured with the prototyping platform is given. The TV signals are generated by a Rohde & Schwarz signal generator. The signals are sent to the prototyping platform for detection. Additionally, the actual signal power is measured using a Rohde & Schwarz power meter. The parameters used for the simulations as well as for the measurements are as follows: The bandwidth used for all TV

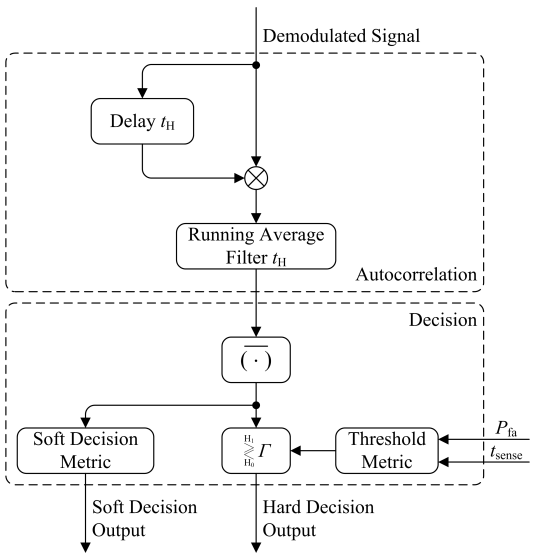


Fig. 10. Flow Diagram of the PAL-D/K Sensing Algorithm

 ISBN: 978-1-61208-203-5

standards is 8 MHz and the sensing interval t_{sense} is set to 20 ms. The target false-alarm probabilities are 10% and 0.1%, while the target detection probability is 90%. For the simulations, a noise figure of 8 dB is considered. Figure 11 shows the detection probability P_{d} versus the received signal power p_{rx}. The considered DTMB signal uses frame header mode 1. The blue curves show the simulation results for a false-alarm probability of 10% and 0.1%, respectively. The red curves show the corresponding measurement results. For the target detection probability of 90%, the measurement results are 3 dB to 4 dB worse than the simulation results. Thus, with the given algorithms, a sensitivity of approximately -110 dBm and -108.5 dBm can be reached in the presented hardware setup. There are several reasons that could result in such degradations. The simulations assume a perfect AGC while in the real system the maximum gain is limited by the tuner module leading to an increased quantization noise in case of very low signal powers at the input of the tuner. The TV tuner shows highly unstable behavior in terms of amplitude and phase when the input signal is very weak. Such a property results from the fact that the tuner is designed for receiving TV signals at significantly higher power levels. The required sensing sensitivity is much higher than the TV receiver sensitivity. This causes unexpected distortion when the signal level is below the target receiver sensitivity. This aspect is also the reason why autocorrelation algorithms are favorable compared to cross-correlation algorithms. Cross-correlation algorithms suffer more seriously from such distortions of the tuner, leading maximally to the same overall performance as the autocorrelation algorithms although in simulations such cross-correlation algorithms perform better than their autocorrelation counterparts. However, the computational complexity of cross-correlation implementations is much higher and, hence, autocorrelation algorithms are preferred.

Further reasons for the difference between the simulated and measured sensing results are effects such as frequency offsets and amplifier non-linearities in the RF stage, which cannot be avoided in hardware implementations and may lead to significant performance degradations. However, these effects have not been considered in the simulations.

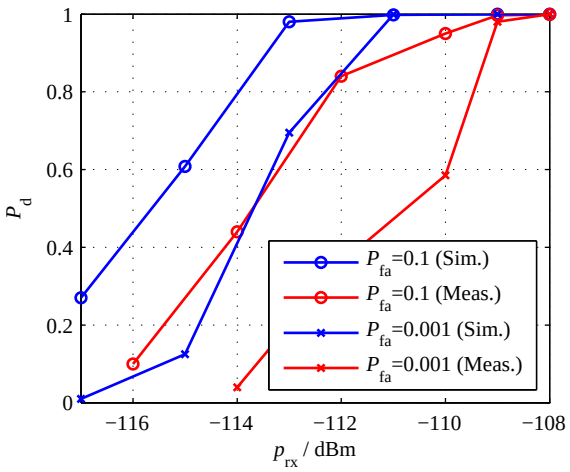


Fig. 12. Simulation and Measurement Results for CMMB

The simulation and measurement results for the TV standard CMMB are plotted in Figure 12. Again, a degradation of the measurement performance of almost 3 dB compared to the simulation performance can be identified. With the implemented algorithms, a sensitivity of -111 dBm ($P_{\text{fa}} = 10\%$) and -109.5 dBm ($P_{\text{fa}} = 0.1\%$), respectively, can be reached for the given target detection probability.

In case of PAL-D/K, the measured performance is much worse than the simulated performance as shown in Figure 13. For PAL-D/K, further signal processing steps are necessary to extract the CVBS signal from the received PAL signal. These signal processing steps need to be carried out before the sensing operation. However, they are implemented in a way to minimize the processing latency rather than for utilizing the dynamic range most efficiently. This leads to a significant

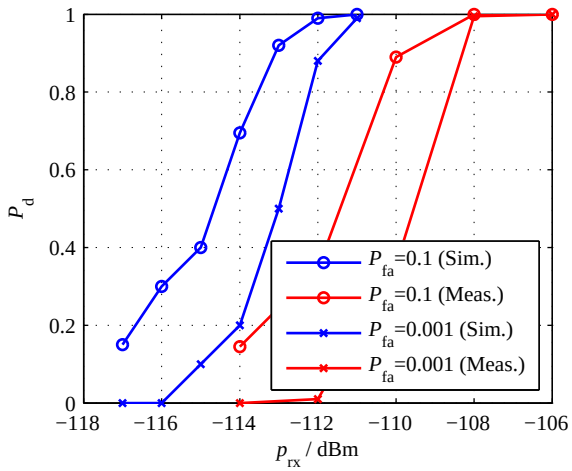


Fig. 11. Simulation and Measurement Results for DTMB

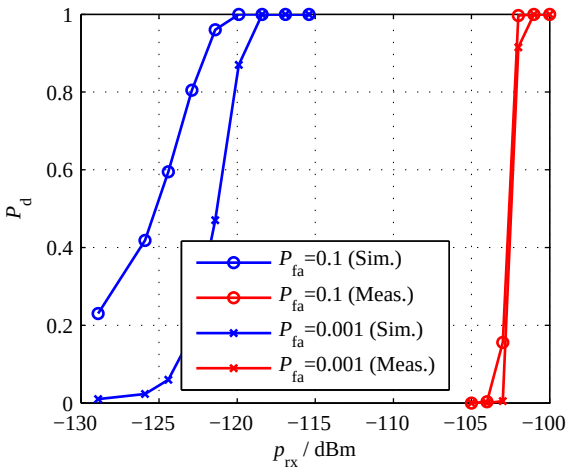


Fig. 13. Simulation and Measurement Results for PAL-D/K

 ISBN: 978-1-61208-203-5

performance degradation in comparison to the simulation results, which are based on floating-point calculations without any constraints regarding the dynamic range. The sensitivity for PAL-D/K is approximately -102 dBm.

VII. Conclusion

In this manuscript, a prototyping platform for spectrum sensing was presented and its underlying architecture was illustrated. Furthermore, an application of cognitive radio for TV white space in China was addressed. Therefore, spectrum sensing algorithms for the three predominant TV standards in China, namely DTMB, CMMB and PAL-D/K, were presented. Those sensing algorithms have been validated on a prototyping platform. The prototyping platform itself as well as the underlying signal flow were highlighted. It was shown that a signal detection even at very low input levels is possible with that platform. For a false-alarm probability of 10% and a detection probability of 90%, a sensitivity of -110 dBm can be achieved for DTMB. For CMMB and for PAL-D/K, -111 dBm and -102 dBm, respectively, can be achieved.

References

[1] J. van de Beek, J. Riihijarvi, A. Achtzehn, and P. Mahonen, "UHF White Space in Europe - A Quantitative Study into the Potential of the 470-790 MHz Band," in *IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, May 2011, pp. 1–9.

[2] FCC, "Report 10-174: In the Matter of Unlicensed Operation in the TV Broadcast Bands, Additional Spectrum for Unlicensed Devices Below 900 MHz and in the 3 GHz Band - Second Memorandum Opinion and Order," Adopted: 23 September 2010.

[3] IEEE, "IEEE 802.22 Working Group on Wireless Regional Area Networks." [Online]. Available: http://www.ieee802.org/22

[4] W. Fischer, *Digital Video and Audio Broadcasting Technology: A Practical Engineering Guide*, 3rd ed. Springer Publishing Company, Inc., 2010.

[5] *Framing Structure, Channel Coding and Modulation for Digital Television Terrestrial Broadcasting System*, Std. GB 20 600-2006, 18 August 2006.

[6] *GY/T 220.1-2006: Mobile Multimedia Broadcasting Part 1 - Frame Structure, Channel Coding and Modulation for Broadcasting Channel*, SARFT - The State Administration of Radio, Film and Television Std. GY/T 220.1-2006, 2006.

[7] *Characteristics of PAL-D Television Broadcasting System*, Std. GB 3174-1995, Adopted: 2 December 1995.

[8] *Standard for Cognitive Wireless Regional Area Networks (RAN) for Operation in TV Band*, IEEE Std. 802.22, July 2011.

[9] FCC, "Evaluation of the Performance of Prototype TV-Band White Space Devices Phase II," 15 October 2008.

[10] V. Gaddam and M. Ghosh, "Robust Sensing of DVB-T Signals," in *IEEE Symposium on New Frontiers in Dynamic Spectrum*, Apr. 2010, pp. 1–8.

[11] A. Xu, Q. Shi, Z. Yang, K. Peng, and J. Song, "Spectrum Sensing for DTMB System Based on PN Cross-Correlation," in *IEEE International Conference on Communications (ICC)*, May 2010, pp. 1–5.

[12] L. Wenqi, X. Ning, G. Lijun, Z. Yingxin, and W. Hong, "Spectrum Sensing Methods for DTMB Based Cognitive Radio Systems," in *1st International Conference on Information Science and Engineering (ICISE)*, Dec. 2009, pp. 2730–2733.

[13] J. Mitola, D. Chester, S. Haruyama, T. Turletti, and W. Tuttlebee, "Globalization of Software Radio," *IEEE Communications Magazine*, vol. 37, no. 2, pp. 82–83, Feb. 1999.

[14] *Digital Video Broadcasting (DVB): Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television (DVB-T)*, ETSI Std. EN 300 744 V1.6.1, January 2009.

 ISBN: 978-1-61208-203-5

Maximal Ratio Combining SC-FDMA Performance over Correlated Ricean Channels

Jyoti R. Gangane
Electronics and Telecommunication Department
Sinhgad Institute of Technology
Lonavala, Pune (India)
ganganejyoti@gmail.com

Mari Carmen Aguayo-Torres, Juan J. Sánchez-Sánchez
Departamento de Ingeniería de Comunicaciones
University of Malaga
Malaga (Spain)
aguayo@ic.uma.es, jjsanch@ic.uma.es

***Abstract*— Long Term Evolution (LTE) system has selected Single Carrier Frequency Division Multiple Access (SC-FDMA) for uplink transmission. Multiple Input Multiple Output (MIMO) can be used in order to improve throughput, robustness, coverage and capacity. Although it is well known that the performance of SC-FDMA is worse than that of Orthogonal Frequency Division Multiplexing (OFDM) for Rayleigh channels, the existence of a Line Of Sight (LOS) is able to increase SC-FDMA resistance to fading further than that of OFDM. In this work, the effect of antenna correlation over SC-FDMA is investigated for Rice and Rayleigh fading channels. Performance of MRC SC-FDMA is compared to that of ZF and MMSE equalizers for several LOS power and antenna correlation values. Results show that SC-FDMA coherent combination performance is better than that of OFDM for both Rayleigh and Rice channels. Influence of fading frequency correlation function on SC-FDMA performance is kept under MRC.**

Keywords-SC-FDMA, MRC, Rice

I. INTRODUCTION

Single Carrier Frequency Division Multiple Access (SC-FDMA) is used for the E-UTRA Long Term Evolution (LTE) mobile communication system. SC-FDMA, also referred to as Discrete Fourier Transform (DFT) spread Orthogonal Frequency Division Multiple Access (OFDM), has been selected for uplink transmission for LTE [1]. The main important advantage of SC-FDMA compared to standard OFDM is its low Peak to Average Power Ratio (PAPR) [2], which enables low complexity implementation of mobile terminal.

In general, it is accepted that SC-FDMA link level performance is worse than that of OFDM [3]. However, under certain conditions, SC-FDMA behavior improves that of OFDM. Specifically, the existence of a Line Of Sight (LOS) increases SC-FDMA resistance to fading over OFDM [4]. Roughly speaking, SC-FDMA BER is obtained from the harmonic average of the channel response at the allocated subcarriers. On the other hand, BER in OFDM is evaluated as the average of the BER for each subcarrier. Under high probability of deeply faded subcarriers (as in Rayleigh channels), SC-FDMA basically behaves as the worst subcarrier. However, the existence of a LOS greatly reduces the probability of deep fading. Without that burden, SC-FDMA is able to add frequency diversity to an OFDM system, thus reducing BER.

In the receiver at base station, frequency domain Zero Forcing (ZF) or Minimum Mean-Squared Error (MMSE) linear equalization [5], [6] might be applied. For OFDM, both techniques obtain similar results [7]. However, MMSE SC-FDMA performance outperforms MMSE OFDM [8][9] due to a similar reason to that of the existence of a LOS: MMSE modifies the harmonic average previously described by including a constant for each subcarrier.

Multiple Input Multiple Output (MIMO) techniques [10] take advantage of the spatial separation between antenna elements to create uncorrelated spatial channels and to exploit higher levels of the spatial diversity. This improves spectral and power efficiency. MIMO techniques are very attractive at base station, where large antenna spacing is easily accommodated. On the mobile unit, however, single antenna is more feasible, thus Single Input Multiple Output (SIMO) techniques are advisable for uplink.

Diversity combining is well known to mitigate the performance degradation in multipath fading. Specifically, Maximal Ratio Combining (MRC) represents an optimal combiner over fading channels: multiple copies of the same information signal are blended so as to maximize the instantaneous SNR at the output [11]. Exact closed form expressions of average symbol error rate (SER) can be found for uncorrelated [12] and correlated [13] received Rayleigh channels, and certain efforts for analysis under LOS reception can also be found [14]. In general, it is well known that antenna correlation degrades the system performance as less diversity is present at the receiver.

MRC for received uplink SC-FDMA signals is implicitly assumed in many works with multiple antennas at the base station [15][6]. However, works regarding evaluation of MRC SC-FDMA link level performance for Rice channels and studies on the effect of antenna correlation are difficult to find.

In this paper, BER performance of MRC SC-FDMA system is evaluated. The effects of antenna correlation and Rice factor on SC-FDMA link level performance are studied and compared to those over Rayleigh channel and for OFDM. Two distinct realistic channel models [16] are used for simulation, as power delay profile greatly influences SC-FDMA performance.

The rest of the paper is organized as follows. In Section II, we summarize the maximal ratio combining receiver diversity scheme. Our system model is described in Section III. In Section IV, we evaluate BER of MRC SC-FDMA system by simulating a BPSK signaling scheme. Finally, some concluding remarks are given.

II. MAXIMAL RATIO COMBINING

MRC is a SIMO technique allowing coherent combination of signals received over a set of antennas. In particular, signals from antenna elements are weighted and combined to maximize the output Signal to Noise Ratio (SNR).

 ISBN: 978-1-61208-203-5

Consider a receiver diversity system with N_R antennas as shown in Fig. 1. The channel can be expressed as

$$\boldsymbol{h} = \left[h_1\ h_2\ h_3\ \ldots h_{N_R}\right]^T \quad (1)$$

The received set of signals $\boldsymbol{y} = \left[y_1\ y_2\ y_3\ \ldots y_{N_R}\right]^T$ is then

$$\boldsymbol{y} = \sqrt{\frac{E_x}{N_0}}\boldsymbol{h}x + \boldsymbol{n} \quad (2)$$

being $\mathbf{n}$ a vector of noise AWGN samples and x the transmitted symbol. Let γ_i be the instantaneous SNR for the i^{th} branch, which is given by

$$\gamma_i = \frac{E_x}{N_o}|h_i|^2 \quad (3)$$

In MRC, all N_R branches are combined by the following weighted sum:

$$y_{MRC} = \left[W_1 W_2 W_3 \ldots W_{N_R}\right]\boldsymbol{y} = \boldsymbol{W}^T\boldsymbol{y} \quad (4)$$

Power of the instantaneous signal and noise part are respectively given as

$$P_{signal} = \frac{E_x}{N_0}\ |\ W^T\boldsymbol{h}|^2 \quad (5)$$

and

$$P_{noise} = \|\boldsymbol{W}^T\|_2{}^2 \quad (6)$$

where $\|\cdot\|_2$ represents the usual Euclidean norm. From equation (5) and (6) the average SNR for MRC is given as

$$SNR_{MRC} = \frac{P_{Signal}}{P_{noise}} = \frac{E_x}{No}.\frac{|\boldsymbol{W}^T\ \boldsymbol{h}|^2}{\|\boldsymbol{W}^T\|_2{}^2} \quad (7)$$

Invoking the Cauchy-Schwartz inequality, SNR is maximized at $\mathbf{W} = \mathbf{h}^*$ which yields

$$SNR_{MRC} = \frac{E_x}{N_0}\|h\|_2{}^2 \quad (8)$$

Roughly, MRC process corrects the channel phase and blends the two received signals in the correct direction. Further, signals are amplitude scaled in such a way that stronger signals are more influent in the final value. Later, the amplitude scaling step makes sure that the received sequence has similar amplitude as the transmitted sequence. These steps together remove the channel effect and replace the equalizer.

III. SYSTEM MODEL

We consider Fig. 2 - Fig. 3 system model with one transmit and N_R receive antenna, i.e. a SIMO $1 \times N_R$ system.

For a given user, a sequence of transmitted bit is mapped to a constellation of complex symbols (e.g. BPSK, QAM). The precoded complex symbol $\mathbf{X}$ is obtained by performing N-DFT operation over the resulting complex sequence $\mathbf{x}$. Then, $\mathbf{X}$ is mapped on a subset of different allocated subcarriers per user,

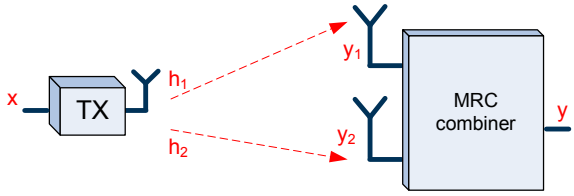


Figure 1. Transmitter and receiver configuration for MRC

i.e., N out of M sub-carriers in which the total system bandwidth is divided. The subset may consist of a group of adjacent localized SC-FDMA (L. SC-FDMA) or of distributed interleaved SC-FDMA (I. SC-FDMA) subcarriers [2]. Non-allocated subcarriers are forced to zero. From this point onwards transmission is similar to that of OFDMA.

The channel system with N_R diversity branches at the receiver can be represented by the channel vector

$$\boldsymbol{h} = \left[h_j\right] \quad (9)$$

where h_j is the channel coefficient between the transmit antenna and the i^{th} receive antenna. Certain correlation, measured through the correlation factor ρ, can exist among paths. Moreover, h_j can be described by multiple paths, which arise from spreading. If there are N_j distinct paths from the transmitter to the receiver, the impulse response for this channel will be:

$$h_j(t,\tau) = \sum_{i}^{N_j} a_i\delta(t,\tau_i) \quad (10)$$

where t stands for time variability and τ for delay. This is the well known tapped-delay line model. Path amplitudes are well described by Rayleigh distributed amplitudes varying according to a classical Doppler spread and with average power as given by the Power Delay Profile (PDP). Moreover, a Line of Sight (LOS) component can also exist between the transmitter and the receiver. The Rice factor K [10] measures the relative strength of the LOS compared to that of the whole varying amplitude. It is a measure of the severity of the fading, being K = 0 the most severe fading case (Rayleigh fading, i.e. no LOS), and K = ∞ the usual Additive White Gaussian Noise (AWGN) channel.

At the receiver, N_R different receiving chains are followed as shown in Fig. 3. Perfect channel estimation and synchronization avoid interference from other users. The cyclic prefix is suppressed and an M-DFT operation converts each time domain symbol in to a frequency-domain symbol at the receiver. After demapping, the received symbols at each antenna $\boldsymbol{Y_n}$ can be expressed as

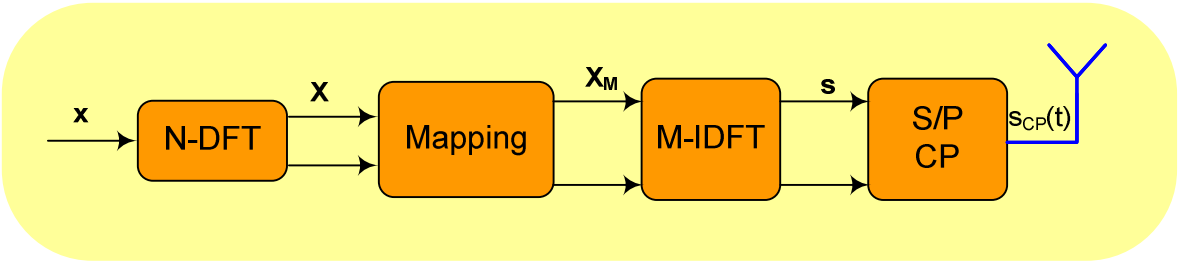


Figure 2. SC-FDMA transmitter scheme

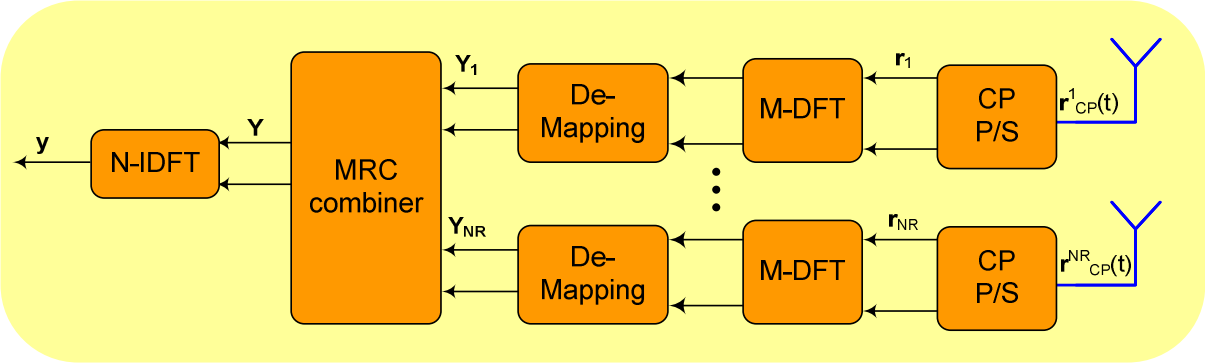


Figure 3. MRC SC-FDMA receiver scheme

$$\boldsymbol{Y}_n = \boldsymbol{H}_n \boldsymbol{X} + \boldsymbol{\eta} \quad (11)$$

where $\boldsymbol{\eta}$ is a noise vector whose entries are i.i.d. complex Gaussian $C\mathcal{N}$ (0, N_0) and $\boldsymbol{H_n}$ represents the $N \times N$ diagonal matrix whose entries are the channel frequency response as seen by antenna n for each allocated subcarrier [10].

These N_R signals are frequency combined using MRC:

$$\boldsymbol{Y} = \frac{\sum_{n=1}^{N_R} \boldsymbol{Y_n}\, \boldsymbol{H_n^*}}{\sum_{n=1}^{N_R} |\boldsymbol{H_n}|^2} \quad (12)$$

After taking IDFT of $\boldsymbol{Y}$, signal is given to detector. The output of detector is the estimated input bit sequence.

IV. Simulation Results

In this section, simulation results are given in order to evaluate BER performance for MRC SC-FDMA. LTE settings are fixed to those in Table I. Channel profiles described in Table II are adopted from ITU-R_M.1225 specs [16] with an added direct LOS with Rice factor K. Following figures give the BER performance of MRC SC-FDMA for two receiving antenna with different antenna correlation ρ, number of allocated subcarrier N and Rice factor K. Results are given for Vehicular B (VB) and Pedestrian A (PA) channels. For comparison, ZF or MMSE frequency equalized single antenna detection [9] are also included in figures, as well as several results for OFDM.

TABLE I. Simulation Parameters

FFT size	2048
Modulation Techniques	BPSK, 16QAM
Carrier frequency	2.00 GHz
System Bandwidth	20 MHz
Sampling Frequency	30.72 MHz
Number of used subcarriers	4, 32
Antenna configuration	$1 \times \{1, 2, 4\}$
Channel model	ITU-R VA & PA channel
Equalizers	MMSE, ZF & MRC
Number of receiving antenna	1, 2, 4

TABLE II. Delay Spread and Coherence Bandwidth for Considered Channels

Channel model	Delay spread (r.m.s.)	Coherence Bandwidth (50%)
PA	46 ns	4.35 MHz
VB	4001 ns	50 KHz

Figs. 4 and 5 show BER of MRC without antenna correlation over VB channel for Localized and Interleaved SC-FDMA, respectively. As it is known, in ZF single antenna reception, OFDMA determines the lower bound for SC-FDMA, while in MMSE. SC-FDMA results are better. It is shown that under MRC, SC-FDMA results are also better than those of OFDM. Improvement for MRC Localized SC-FDMA is slightly higher than that of Interleaved SC-FDMA.

The effect of channel correlation among antenna can be inspected in Fig. 6. Results for Interleaved SC-FDMA over channel were as expected: the more antenna correlation, the more errors at detection. However, performance under Rice fading correlation is less affected by correlation value as direct path affect both antenna anyway. Due to the same reason, specific frequency correlation function influences less SC-FDMA performance under a LOS [9]. However, Fig. 7 shows that VB channel results are better than those of PA channel for MRC Localized SC-FDMA.

Fig. 8 gives results equivalent to those of Figs. 6 and 7 but for OFDM. Note that no effect due to PA or VB channel model or localized/interleaved mode exists in OFDM. Effect of correlation factor is more noticeable than in SC-FDMA. Expected 3dB gain for Rayleigh coherent MRC combination ($\rho = 1$) can be found.

In Fig. 9, results for VB and PA channels are given for two different numbers of allocated subcarriers in Localized and Interleaved SC-FDMA. Over VB channel, improvement is better for a higher number of subcarriers as the probability of at least one very faded pair of carriers is lower. In general, PA channel performance is worse as its coherence bandwidth is higher.

V. Conclusion

In this paper, we have investigated BER performance of MRC SC-FDMA over Rayleigh and Rice fading channels. It is known that ZF SC-FDMA behavior is worse than that of OFDM for single antenna systems. However, coherent combination of signal received on two antennas improves SC-FDMA up to overtake OFDM. Influence of fading frequency correlation function on SC FDMA performance is kept under MRC. Lower correlation among allocated

 ISBN: 978-1-61208-203-5

subcarriers (i.e., higher frequency diversity) improves performance of SC-FDMA (VB channels vs. PA fading; Interleaved mode vs. Localized).

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support of the European Commission through Erasmus Mundus External Cooperation Window "Mobility for Life" and express their gratitude to Junta de Andalucía (Proyecto de Excelencia P07-TIC-03226), to the Spanish Government (Plan Nacional de I+D+I, TEC2010-18451), and to Sinhgad Technical Education Society (India).

REFERENCES

[1] 3GPP TS 36.211, "Evolved universal terrestrial radio access (E-UTRA): physical channels and modulation (Release 8)," June 2007

[2] H. Myung, J. Lim, and D. Goodman, "Single carrier FDMA for uplink wireless transmission," IEEE Veh. Tech. Mag., vol. 1, pp.30-38, Sep. 2006

[3] J. Zhang, C. Huang, G. Liu, and P. Zhang, "Comparison of the Link Level Performance between OFDMA and SC-FDMA," Proc. of First Int. Conf. Comm. and Networking in China, ChinaCom'06, Oct. 2006, pp.1-6, 25-27

[4] J.J. Sánchez-Sánchez, U. Fernández-Plazaola, and M.C. Aguayo-Torres, "BER analysis for zero-forcing SC-FDMA over Nakagami-m fading channel," IEEE Trans. Veh. Tech., vol. 60, no. 8, 4077 – 4081, Oct. 2011

[5] T. Lunttila, J. Lindholm, K. Pajukoski, E. Tiirola, and A. Toskala, "EUTRAN uplink performance," Proc. Int. Symp. Wireless Pervasive Computing (ISWC'07), San Juan, Puerto Rico, pp. 515-519, Jan. 2007

[6] B. Priyanto, H. Codina, S. Rene. T. Sorensen, and P. Mogensen, "Initial performance evaluation of DFT –spread OFDM based SC-FDMA for UTRA LTE uplink," Proc. of IEEE Veh. Tech. Conf. (VTC-Spring 2007), Dublin, Ireland, Apr. 2007, pp. 3175-3179

[7] Z. Wang, X. Ma, and G. Giannakis, "OFDM or single-carrier block transmissions?", IEEE Trans. Comm., vol. 52, no. 3, pp. 380 - 394, March 2004

[8] C. Ciochina, D. Castelain, D. Mottier, and H. Sari, "Single carrier space frequency block coding: performance evaluation," IEEE Veh. Tech. Conf., pp. 715-717, Oct. 2007

[9] J.R. Gangane, M.C. Aguayo-Torres, J.J. Sánchez-Sánchez, and U. Fernández-Plazaola, "SC-FDMA performance over Ricean channels," Proc. of IEEE Int. Biomedical and Broadband Comm., IB2Com 2011, Melbourne, Australia, Dec. 2011

[10] A.J. Goldsmith, Wireless Communications, Cambridge University Press, 2005

[11] D. Brennan, "Linear diversity combining techniques," Proc. IRE, vol. 47, no. 1, pp. 1075–1102, June 1959

[12] J. Lu, T.T. Tjhung, and C.C. Chai, "Error Probability Performance of –Branch Diversity Reception of MQAM in Rayleigh Fading", IEEE Trans. Comm., vol. 46, no. 2, Feb. 1998

[13] L. Fang, G. Bi, and A.C. Kot, "New Method of Performance Analysis for Diversity Reception with Correlated Rayleigh-fading Signals", IEEE Trans. Veh. Tech., vol. 49, No. 5, Sept. 2000

[14] L. Najafizadeh and C. Tellambura, "BER Analysis of Arbitrary QAM for MRC Diversity With Imperfect Channel Estimation in Generalized Ricean Fading Channels," IEEE Trans. Veh. Tech., vol. 55, no. 4, pp.1239-1248, July 2006

[15] S. Sesia, I. Toufik, and M. Baker, "LTE – The UMTS Long Term Evolution. From Theory to Practice", Wiley, 2009

[16] International Telecommunication Union (ITU), "Recommendation ITU-R M-1255; Guidelines for evaluation of radio transmission technologies for IMT-2000," Tech. Rep. 1997

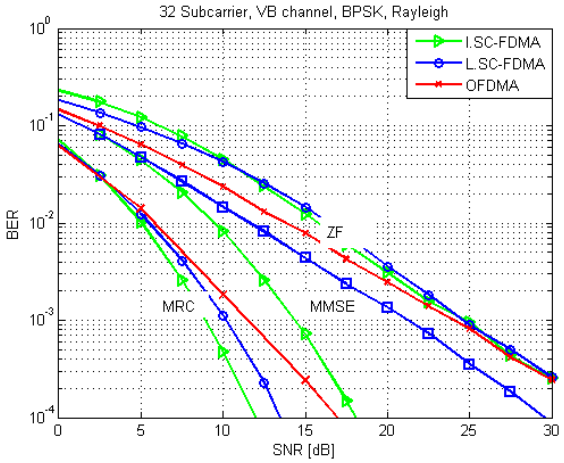


Figure 4. BER of MRC SC-FDMA over Rayleigh fading VB channel

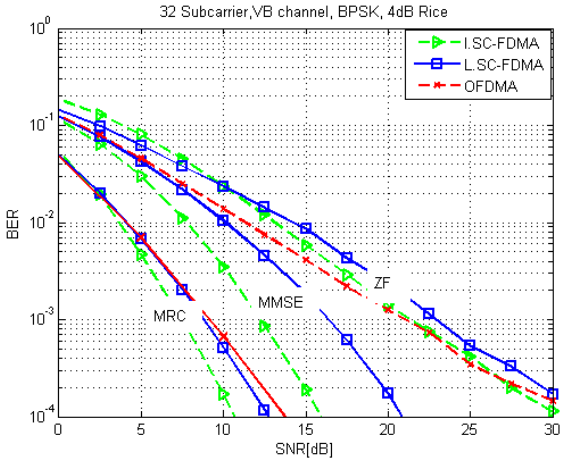


Figure 5. BER of MRC SC-FDMA over Rice fading VB channel

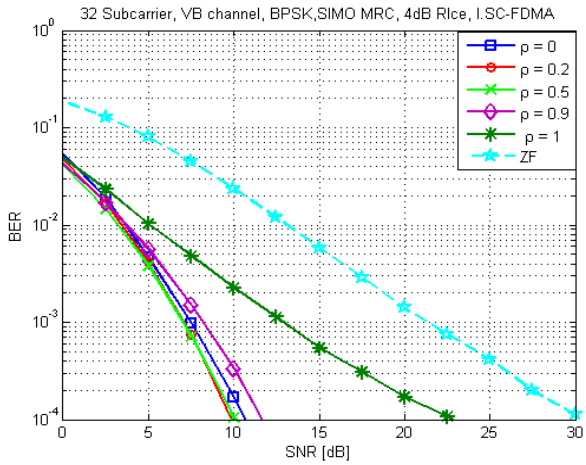


Figure 6. BER of Interleaved SC-FDMA with MRC for different antenna correlation factor over Rice fading VB channel

 ISBN: 978-1-61208-203-5

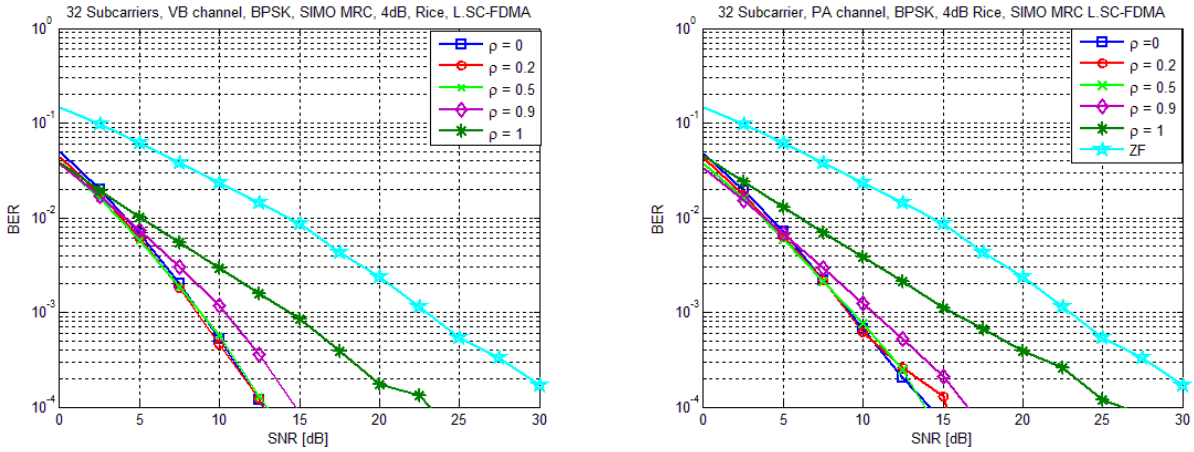


Figure 7. BER performance of localized VB (left) and PA (right) SC-FDMA with MRC for different antenna correlation factor over Rice fading

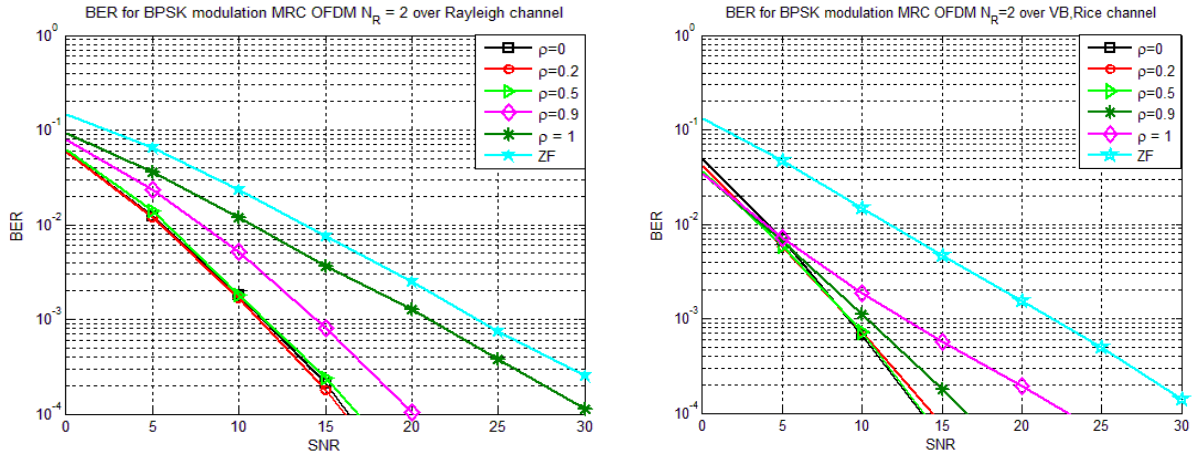


Figure 8. BER of OFDM with MRC for different antenna correlation factor over Rayleigh (left) and Rice 4dB (right) fading

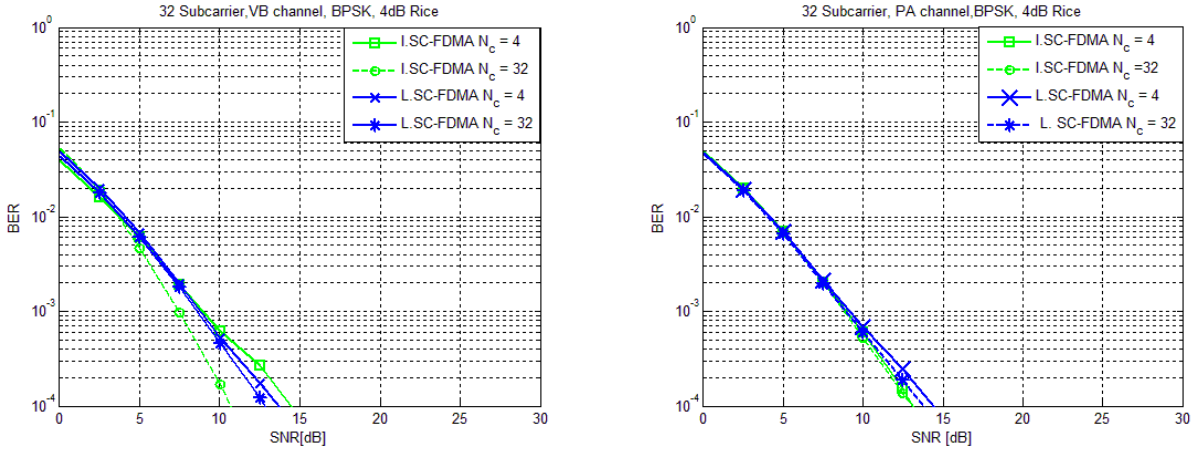


Figure 9. BER performance of SC-FDMA with MRC for different number of allocated subcarriers over VB and PA Rice fading channels

 ISBN: 978-1-61208-203-5

Distribution of 2.4 GHz Range Radiowaves Indoors

Alexey Lagunov
Department of Computer Science and Electronic Devices
North (Arctic) Federal University
named after M.V. Lomonosov
Archangelsk, Russian Federation
a.lagunov@narfu.ru

Darina Lagunova
Department of Computer Science and Electronic Devices
North (Arctic) Federal University
named after M.V. Lomonosov
Archangelsk, Russian Federation
d.lagunova@narfu.ru

***Abstract*— The article reports the results of the research on reducing handicaps level to radio signal in a Wi-Fi network. The authors consider the theory of multi-media in order to understand the processes taking place during reflection of electromagnetic waves with a frequency of 2.4GHz. The resulting numerical modeling conclusions are used to develop measures for the processing of premises multilayer materials. Experiments have shown that the rate of data transmission in wireless IEEE 802.11n standard after treatment of premises increased by 15-20%.**

Keywords - RadioEthernet; wireless interference; disturbance; Wi-Fi, reflectance.

I. INTRODUCTION

Recently, content of data transmitted has more changed to the side of multimedia. This leads to an increase in the volume of data transmitted. To transmit large volumes of data need to increase the data rate. Adopted in 2009, the IEEE 802.11n declares transmit rate of 300 Mbit/s, but the real data transmit rate is 20-30% of the declared. The statistical theory of radiowave distribution indoor is described in our paper [1]. It offers a way to increase the speed of the network. Let us consider another way.

Section "Interference handicaps" is devoted to research features the work of the interference noises of wireless networks. In the section "Definition of factor of reflection interference materials», we consider the theory of the behavior of the reflection coefficient of the vertical and horizontal polarization plane waves at oblique incidence in the controlled environment. The section "Definition of permittivity" is devoted to research of one of the methods for determining the different materials dielectric permittivity ε *. On the basis of the developed theory, we conducted a pilot research that is presented in section "Application of the geometric theory at construction of Wireless Networks".

II. INTERFERENCE HANDICAPS

Interference handicaps, arising due to repeated radiowaves reflection from surrounding subjects, are shown in simultaneous receipt in the receiver of useful signal "copies" set with the displaced phases that can result in its easing or even full disappearance on separate sites of the spectrum (so-called "fading").

Under the same system Direct Sequence Spread Spectrum (DSSS) external conditions appear steadier to fading, than Frequency Hopping Spread Spectrum(FHSS) (as well as in case of the narrow-band handicaps, the useful signal appears deformed only on separate frequencies); however, they are much more sensitive to displacement in time of the protected binary signal - because of considerably shorter (approximately ten times) pulses duration the levels wrong interpretation probability 0 or 1 grows at gate.

At electromagnetic radiation interaction with materials in the last absorption (dielectric and magnetic decreasees), dispersion (due to structural heterogeneity of a material) and radiowaves interference take place. Non-magnetic materials from the radio signal absorption view subdivide on interference, gradient and combined. Interference materials will consist of alternating dielectric and conducting layers. The waves reflected from electro conductive layers and from a protected object metal surface interfere among themselves in them. Gradient materials (the most extensive class) have multilayered structure with smooth or step change of complex dielectric permeability on thickness (it is usual under the hyperbolic law). Their thickness is rather great and makes> 0.12-0.15 λmax, where λmax - the maximal working wave length (in our case 0.12м). The external (matching) layer is made from firm dielectric with the big maintenance of air inclusions, with the dielectric permeability close to unit, with other (absorbing) layers - from dielectric with high dielectric permeability with absorbing conducting stuff. Also materials with a relief external surface (formed by ledges as thorns, cones and pyramids), named subulate materials are conditionally related to gradient materials. Reflection's factor reduction is promoted by repeated waves reflection from thorns surfaces (with waves energy absorption at each reflection) in them. The combined materials - a combination of gradient and interference materials. They differ in action efficiency in the expanded wave band.

The greatest level interfering handicaps is provided with signals at direct falling a radiowave on a material. At application of not directed aerial access point is a wall or a ceiling to which the aerial fastens. Application gradient material for processing a wall or a ceiling in the aerial fastening point can provide increase in a ratio signal/noise up to 6 dB. The aerial direction on a concave surface is inadmissible, as it results in a high level interfering

 ISBN: 978-1-61208-203-5

handicaps and high non-uniformity of a radio signal. Walls on which the direct radiowave gets are processed interfering or gradient materials.

III. DEFINITION OF FACTOR OF REFLECTION INTERFERENCE MATERIALS

The most difficult for practical work is reflection factor definition of materials premises used for processing. Radiophysical diagnostics systems work and the control interference environments are based on the reaction analysis of the researched environment on probing signal.

One of the most actual the problem is problem of the electromagnetic waves interaction adequate description with sound the environment characterized by complex dielectric permeability (ε *) by the sounding data interpretation methods development. It is connected by that sound material environments represent complex dielectric structures. These environments constantly contact to a variable temperature field and water in its various modular conditions in real conditions. These variable components define, basically, dielectric properties of such environments.

It is necessary to take into account spatial distribution ε * at the of radiowave diagnostics of a condition and properties of such environments problems decision. The data about profile distribution ε * can be received or from the aprioristic data, or using the approached theoretical models, or experimentally. One of the reflected signals interpretation methods perfection directions is connected with the modeling tasks decision of which are taking into account flat waves interaction with the layered environment which is described by real geometrical parameters and real dielectric characteristics.

Let us analyze reflection factor behaviour of flat waves of vertical and horizontal polarization at inclined falling on the controllable environment. Sharing of vertically and horizontal the polarized waves results allow to take the information on dielectric properties layers.

Statement of a task

On the flaky-non-uniform dielectric environment from free space (ε * =*1*, μ * = *1*) the flat electromagnetic wave under various Θ angels (Fig. 1) falls.

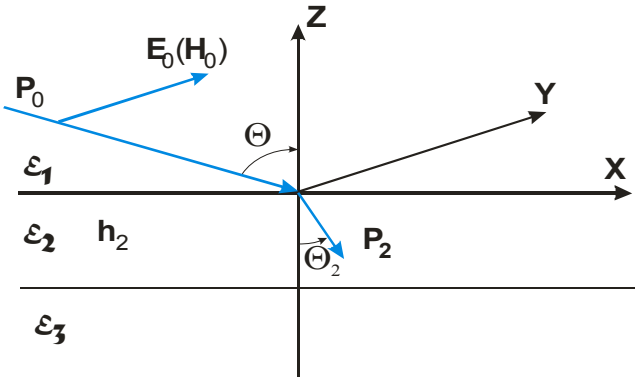


Figure 1. Geometry of a problem

It is required to define values of factor of reflection K from the researched environment depending on a horizontal and vertical polarization waves falling angel in case of a presence on the dielectric layers environment surface. The top and bottom layers have half-infinite thickness, and thickness of second rather thin layer is a variable quantity and commensurable with a wave length. Values ε * the second and third layer change during experiment.

Physical model

Environment with profile distribution ε usually is represented as multilayered system for carrying out of numerical modeling. In this case, ε * will be function of coordinate *Z*, and on borders between layers this function can be explosive. Dependence ε * *(Z)* inside each layer is set by numerical values in some points *Zi*. We consider ε * between points *Zi* and *Zi+1* constant and homogeneous in *X* and *Y* directions on layers for simplification of calculations.

Mathematical model

The multilayered environment reflection factor is defined under the recurrent formula [2]:

$$K_{1,n} = \frac{K_{1,2} + K_{2,n} e^{-j\frac{4\pi h_2}{\lambda\sqrt{\varepsilon_2}}}}{1 + K_{1,2} K_{2,n} e^{-j\frac{4\pi h_2}{\lambda\sqrt{\varepsilon_2}}}} \quad (1)$$

$$K_{i,i} = 0, \quad K_{i,i+1} = \frac{\sqrt{\varepsilon_{i+1}} - \sqrt{\varepsilon_i}}{\sqrt{\varepsilon_{i+1}} + \sqrt{\varepsilon_i}} \quad (2)$$

$$K_{i,k} = \frac{K_{i,i+1} + K_{i+1,k} e^{-j\frac{4\pi h_{i+1}}{\lambda\sqrt{\varepsilon_{i+1}}}}}{1 + K_{i,i+1} K_{i+1,k} e^{-j\frac{4\pi h_{i+1}}{\lambda\sqrt{\varepsilon_{i+1}}}}}, \quad k \neq i, k \neq i+1 \quad (3)$$

Using formulas (1-3), we will find formulas for reflection $K_{1,3}$ factor in case of the research model accepted by us:

$$K_{1,3} = \frac{K_{1,2} + K_{2,3} e^{\gamma_1}}{1 + K_{1,2} K_{2,3} e^{\gamma_1}}, \quad \text{де} \quad \gamma_1 = -j\frac{4\pi h_2}{\lambda\sqrt{\varepsilon_2}} \quad (4)$$

Then for horizontal polarization:

$$K_{1,2} = \frac{\sqrt{\varepsilon_1}\cos\Theta - \sqrt{\varepsilon_2 - \varepsilon_1(\sin\Theta)^2}}{\sqrt{\varepsilon_1}\cos\Theta + \sqrt{\varepsilon_2 - \varepsilon_1(\sin\Theta)^2}} \quad (5)$$

$$K_{2,3} = \frac{\sqrt{\varepsilon_2}\cos\Theta_2 - \sqrt{\varepsilon_3 - \varepsilon_2(\sin\Theta_2)^2}}{\sqrt{\varepsilon_2}\cos\Theta_2 + \sqrt{\varepsilon_3 - \varepsilon_2(\sin\Theta_2)^2}}, \quad (6)$$

$$\text{где } \Theta_2 = \arcsin\left(\frac{\sin\Theta}{\sqrt{\varepsilon_2}}\right)$$

For vertical polarization:

$$K_{1,2} = \frac{\varepsilon_2\cos\Theta - \sqrt{\varepsilon_1(\varepsilon_2 - \varepsilon_1(\sin\Theta)^2)}}{\varepsilon_2\cos\Theta + \sqrt{\varepsilon_1(\varepsilon_2 - \varepsilon_1(\sin\Theta)^2)}} \quad (7)$$

$$K_{2,3} = \frac{\varepsilon_3\cos\Theta_2 - \sqrt{\varepsilon_2(\varepsilon_3 - \varepsilon_2(\sin\Theta_2)^2)}}{\varepsilon_3\cos\Theta_2 + \sqrt{\varepsilon_2(\varepsilon_3 - \varepsilon_2(\sin\Theta_2)^2)}} \quad (8)$$

Reflection factors modules for wave's horizontal |*KH*| and vertical |*KV*| polarization have been designed at various falling *Θ* angels on sound environment by formulas (4−8) for different environment conditions (Fig. 1). Thus thickness of a thin layer *h2* varied, various values *ε* * a thin layer and the third layer were set. For presentation thickness of a thin layer was set in relative units and normalized thus to a wave length in the environment

$$\gamma_1 = -j\frac{4\pi h_2}{\lambda\sqrt{\varepsilon_2}} = -j\frac{4\pi H}{\varepsilon_2}, \quad \text{где } H = \frac{h_2}{\lambda_{avg}}, \lambda_{avg} = \frac{\lambda}{\sqrt{\varepsilon_2}}$$

The formulas (4-8) analysis and diagrams on Figs. 2-4 shows, that factors of reflection |*KH*| and |*KV*| on Figs. 2-3 behave classically, as in case of falling a flat wave on the homogeneous dielectric environment. Diagrams |*KH*| monotonously grow from the minimal value at *Θ = 0* up to maximal - at *Θ =90*. Dependence |KV| from a falling angel has more complex kind. In the beginning of coordinates diagrams monotonously decrease up to zero, and then grow up to unit more sharply. Position of a minimum on the diagram depends on thickness and *ε* * thin a layer, and also ε * the third layer (Figs. 2-7). Besides concurrence of diagrams |*KV*| and |*KH*|, received is observed in case of a thin layer absence, with diagrams when a thin layer thickness is equal *0.5λε* (Figs. 2,3-6,7). This fact indicates that the reflected waves from a thin layer and environment are summarized in a phase.

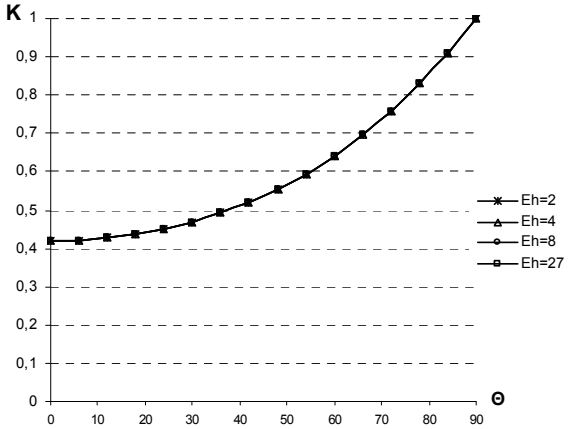


Figure 2. $K = \varphi(\Theta)$ (H=0, horizontal polarization)

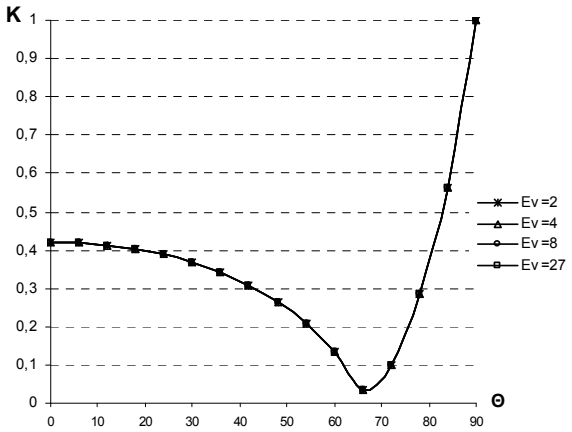


Figure 3. $K = \varphi(\Theta)$ (H=0, vertical polarization)

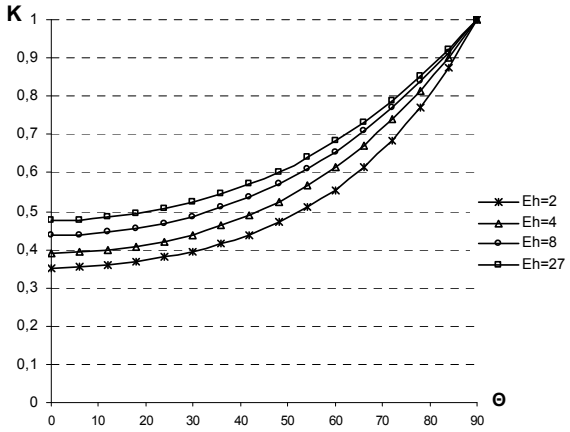


Figure 4. $K = \varphi(\Theta)$ (H=0.25, horizontal polarization)

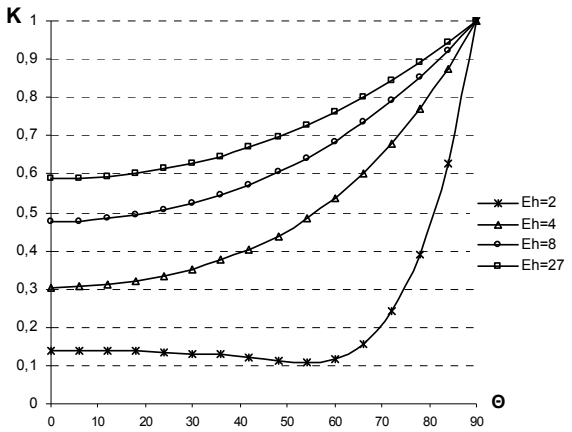


Figure 5. $K = \varphi(\Theta)$ (H=0.25, vertical polarization)

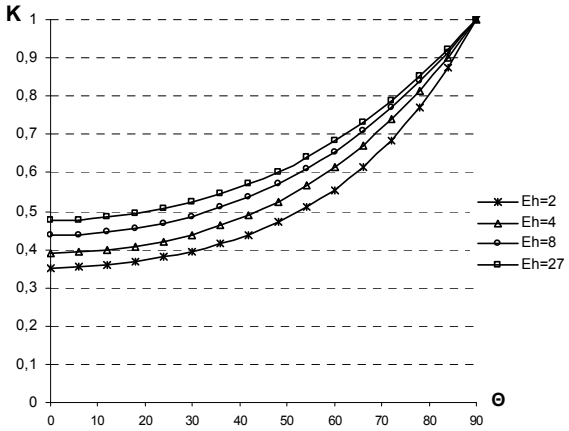


Figure 6. $K = \varphi(\Theta)$ (H=0.5, horizontal polarization)

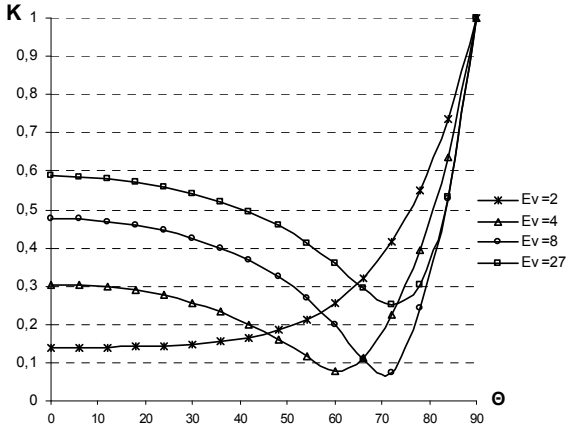


Figure 7. $K = \varphi(\Theta)$ (H=0.5, vertical polarization)

In certain situations the behavior of diagrams |*KV*| and |KH| differs from considered above (Figs. 8-9). At the certain values *ε1, ε2, ε3* layers and thickness of a thin layer equal *0.25λε* the reflected waves from the top layer and a spreading surface are summarized in an antiphase, as results in change of a kind of diagrams |*KV*| and |*KH*|.

For reflection |*KV*| and |*KH*| factors behavior presentation from a falling angel *Θ* and thin layer *H* thickness are constructed three-dimensional diagrams (Figs. 8-15). Value $\varepsilon_3 = 6 - 0.1i$ was supported to constants, value ε2 changed in limits from $2 - 0.1i$ up to $27 - 0.1i$. Thin layer *h2* thickness changed from 0 up to *λmax*.

The figures analysis confirms characteristic failures presence on diagrams |*KV*| and |*KH*| which appear at certain parities *ε1, ε2, ε3* layers, a falling angel *Θ* and a thin layer thickness *N*. Depth of failures on diagrams depends on presence of decreases in the environment and a thin layer. The fact of presence of special points in behavior of factors of reflection |*KV*| and |*KH*| can be used for development of algorithms of definition *ε* * or thickness of a thin layer.

Presence of characteristic failures on diagrams at small values *ε2* allows picking up such material, the reflection factor from which will be minimal (Figs. 8-11). At the big size *ε2* characteristic, dependence is observed; the increase in a thin layer *h2* thickness results in increase in reflection factor (Figs. 12-15).

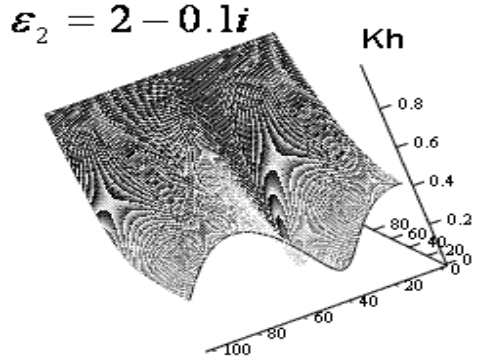


Figure 8. $K = \varphi(\Theta)$

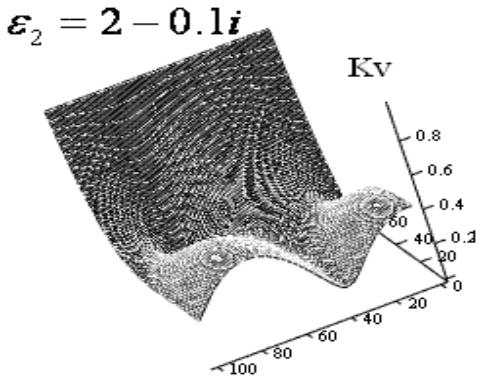


Figure 9. $K = \varphi(\Theta)$

 ISBN: 978-1-61208-203-5

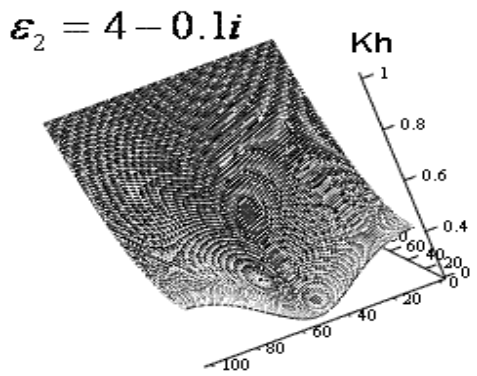


Figure 10. $K = \varphi(\Theta)$

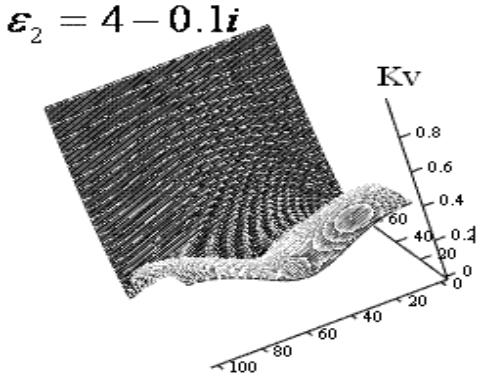


Figure 11. $K = \varphi(\Theta)$

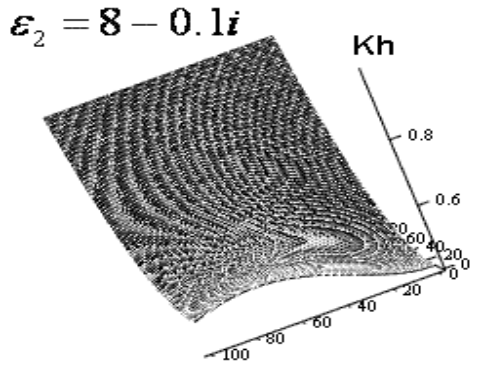


Figure 12. $K = \varphi(\Theta)$

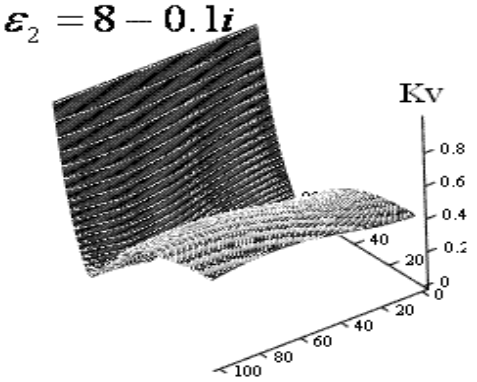


Figure 13. $K = \varphi(\Theta)$

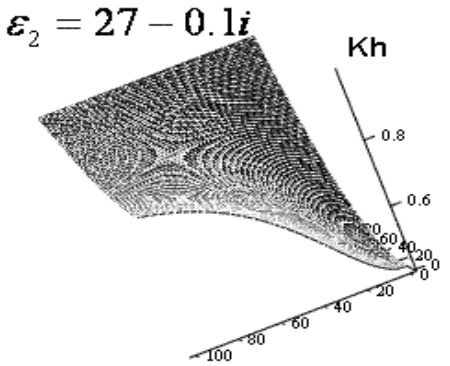


Figure 14. $K = \varphi(\Theta)$

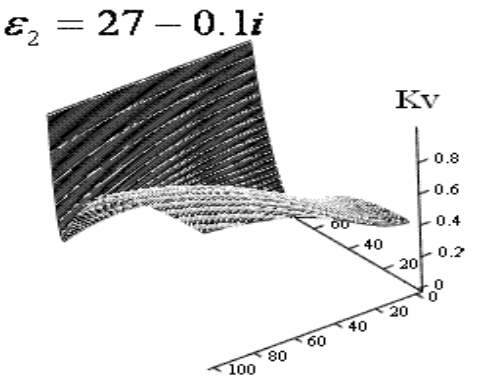


Figure 15. $K = \varphi(\Theta)$

IV. DEFINITION OF PERMITTIVITY

For reflection coefficient definition method use is necessary to know size of permittivity $\varepsilon^* a$ material. For the majority of the materials used in premises furnish, the given value is unknown and there is a permittivity value definition problem.

 ISBN: 978-1-61208-203-5

We research theoretically an opportunity of application of linear aerials for measurement of thickness (*h2*) and permittivity (*ε2*) first layers of the two-layer environment on a variable frequency method. Let us define *h2* and *ε2* by measurement results of an ultrahigh-frequency linear aerials entrance impedance available above environment in turn. Linear aerials impedance measurements are carried out with the help of a transfer complex factors measuring instrument.

Let us assume, that *aerial A* in length *2l* is set at height *h* above the horizontal - layered environment in parallel a surface of environment (Fig. 16). Environment consists of two layers. The first layer is characterized by thickness *h2* and complex permittivity *ε2 **, the second layer - thickness h3 and complex permittivity *ε3**.

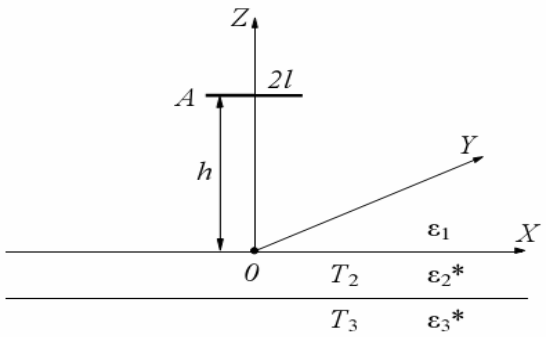


Figure 16. The plan to a problem about definition of an impedance of the linear aerial located above the two-layer environment

Assume that the environment first layer is dielectric, and the second layer - conductive. The first layer thickness is finite, the second layer represents half-subspace ($h3 \to \infty$). Let us consider three cases. In the first case the first layer will be a pine board (*ε2 = 2.73, h2 = 0.07 м*), in the second case - a burnt brick (*ε2 = 5.5, h2 = 0.066 м*), in the third case - the block from a glass (*ε2 = 6, h2 = 0.117 м*). In all three cases value of the factor of decreases of the first layer we shall accept equal 0.01. Let aerials will be adjusted on frequencies 300, 350.. 2200 MHz (with step 50 MHz).

The length of each aerial without taking into account the aerial thickness is determined n the following way:

$$2l = \frac{\lambda}{2} = \frac{c}{2f}, \qquad (9)$$

where

λ – length of a wave in free space, *m*,

c – speed of distribution of waves in free space, *m/s*

f – frequency of tuning of the aerial, *Hz*.

Each aerial is above environment at the height equal to optimum height of the half-wave linear aerial arrangement above a homogeneous environment $h = 0.28\lambda$.

Under condition of an the half-wave linear aerial arrangement above a homogeneous environment at the height equal or not enough distinguished from 0.28λ, the maximal value of the module of an impedance of the aerial is observed.

Results of the calculation executed with use of theoretical model [3], are submitted in Fig. 17 as diagrams of dependences of the half-wave linear aerial impedance module located above the two-layer environment, from the aerial tuning frequency. The curve 1 conforms to a case when the environment first layer is the pine board, curve 2 - a burnt brick, a curve 3 - the block from a glass.

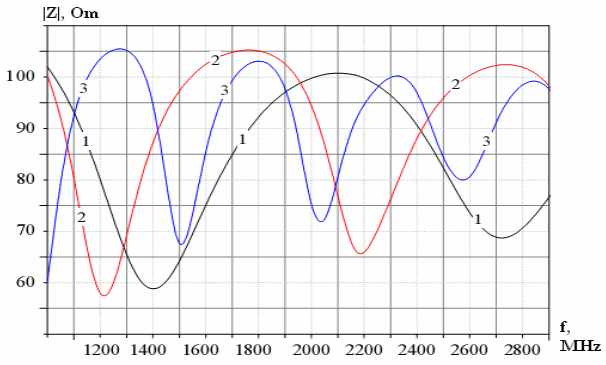


Figure 17. Dependence of the module of an the half-wave linear aerial impedance located above the two-layer environment, on frequency of tunnig at various characteristics of the first layer

1: $\varepsilon 2^* = 2.73 - 0.01i$ (pine board), $h2 = 0.07$ *m*;
2: $\varepsilon 2^* = 5.5 - 0.01i$ (burnt brick), $h2 = 0.066$ *m*;
3: $\varepsilon 2^* = 5.3 - 0.0035i$ (glass), $h2 = 0.12$ *m*

Under the diagrams submitted in Fig. 17, it is possible to determine one of first layer parameters of the two-layer environment (*h2* or *ε2*) if another is known. For calculation of thickness of the first layer we shall use the formula

$$h_2 = \frac{c}{4\sqrt{\varepsilon_2}\cdot\Delta f}, \qquad (10)$$

where Δf - *A* difference of the frequencies corresponding to two next minimum of frequency dependence of the module of an impedance of the linear aerial, Hz,

$$\Delta f = \frac{\left|f_{\min 2} - f_{\min 1}\right|}{2}, \qquad (11))$$

For example, for a case when the environment first layer is the pine board; on a curve 1 in Fig. 17, we find *fmin1* = 1401.542 MHz and *fmin2* = 2711.424 MHz.

Thus Δf = 654.941 MHz. Having substituted values *ε2* and Δf in the formula (10), we receive value *h2*, equal 0.069 m. In this case, the deviation of settlement value of thickness of the first layer from a preset value *(Δh2)* is equal 1.4%.

 ISBN: 978-1-61208-203-5

Similarly we determine a material thickness for a burnt brick and glasses. We use the minimal values close to frequency researched by us 2.4GHz.

Results of calculation under the formula (10) the two-layer environment first layer thickness of are Table I. In the considered cases the deviation of settlement value *h2* from a preset value does not exceed 2%.

TABLE I.
RESULTS OF CALCULATION OF THICKNESS OF THE FIRST LAYER OF THE TWO-LAYER ENVIRONMENT

Materials	**Set point h2, m**	**Design value h2, m**	**Δh2, %**
Pine board	0.07	0.069	1.4
Burnt brick	0.066	0.067	1.50
Glass	0.12	0.122	1.7

For calculation of the two-layer environment first layer permittivity we use the formula

$$\varepsilon_2 = \left[\frac{c}{4h_2\Delta f}\right]^2 \quad (12)$$

Let us define size of dielectric permeability for three materials at known thickness of the first thin layer.

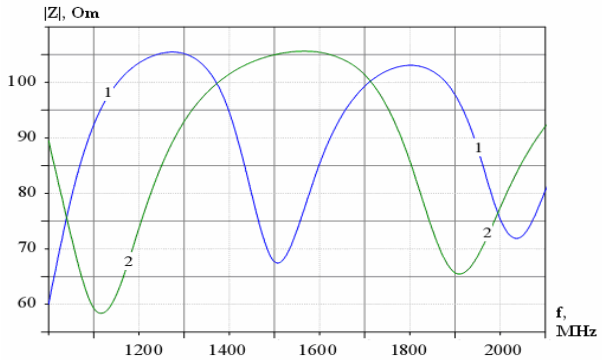
Results of calculation under the formula (12) permittivity of the first layer of the two-layer environment are Table II. In the considered cases the deviation of settlement value *ε2* from a preset value does not exceed 3%.

TABLE II.
RESULTS OF CALCULATION OF PERMITTIVITY OF THE FIRST LAYER OF THE TWO-LAYER ENVIRONMENT

Materials	**Set point ε2,**	**Design value ε2,**	**Δ ε2, %**
Pine board	2.73	2.68	1.8
Burnt brick	5.5	5,58	1.5
Glass	5.3	5.449	2.8

For the definition of permittivity dielectric, which thickness is unknown, the following algorithm is applied. First the difference of frequencies Δ*f1* is defined, corresponding to unknown thickness of a layer *h2* then the superficial part of a layer having thickness Δ*h2* is removed. Then the difference of frequencies Δ*f2* is defined, corresponding to the stayed thickness of a layer (*h2* - Δ*h2*), and depend on *ε2*. For research of a permittivity dielectric measurement opportunity with unknown thickness we return to third of the considered cases.

Let us reduce thickness of the first layer (the block from a glass) by 0.036 *m* and we calculate the module of an linear aerials impedance, serially available above the two-layer environment.



1: $\varepsilon 2^* = 5.3 - 0.035i$ (glass), $h2 = 0.12$ *m*;
2: $\varepsilon 2^* = 5.3 - 0.035i$ (glass), $h2 = 0.084$ *m*

Figure 18. Dependence of the module of an impedance of the half-wave linear aerial located above the two-layer environment, on frequency of tuning at various thickness of the first layer

Results of calculation are submitted in Fig. 18 as a curve of 2 the half-wave linear aerial impedance module dependences located above the two-layer environment, from frequency of tunnig of the aerial. The curve 1 corresponds to a case when a glass layer thickness is equal 0.12 m, and is a part of the curve 3 represented in Fig. 17.

Under the diagrams submitted in Fig. 18, it is possible to define *ε2*, not knowing *h2*. For calculation of the first layer permittivity when its thickness is unknown, we shall use the formula

$$\varepsilon_2 = \left[\frac{c(\Delta f_2 - \Delta f_1)}{4\Delta h_2 \Delta f_1 \Delta f_2}\right]^2 \quad (13)$$

where
Δ*f1* – The difference of frequencies corresponding to thickness of the first layer *h2*, Гц,
Δ*f2* – The difference of frequencies corresponding to thickness of the first layer (*h2* - Δ*h2*), Hz,
Δ*h2* - a difference of thickness of the first layer, m.

Having substituted in the formula (13) values Δ*f1* and Δ*f2*, found under the diagrams represented in Fig. 18, we receive the value *ε2* equal 5.256. In this case, the deviation of settlement value *ε2* from a preset value is equal 1 %.

V. APPLICATION OF THE GEOMETRIC THEORY AT CONSTRUCTION OF WIRELESS NETWORKS

Having received theoretical calculation results the wireless network practical research in indoor is carried out. For treatment premises, we used the multi-layered materials, combining materials with high and low dielectric constant. We used the wireless network Radio Ethernet, making with

 ISBN: 978-1-61208-203-5

standard equipment IEEE 802.11n usage. Router Linksys WRT610N, Netgear WNDR3700 and TRENDnet TEW-671BR are used as POP.

The research was carried out on the basis of method [4] and rate was measured by IxChariot [5].

TCP-traffic (with max size package mainly) is generated by the program and different situation as receiving, transmission and both synchronous (direction to adapter in PC) is modeling. POP (Depending from model no all point was available) is set to operate with 802.11n range on channel 1(5) in regime «40 MHz», previous generation network security regime was switched off, ciphering WPA2-PSK whit c AES algorithm was switched on. Other settings were standard.

That network works sufficiently stable should take into account, as data transmission rate negligible changed during all test. After the first cycle of measurements, we treated the room multilayer materials. Special attention was given to surface where the falling electromagnetic wave at the first reflection. Then there was held the measurements second series.

Test results are shown on Figs. 19, 20, 21, and 22.

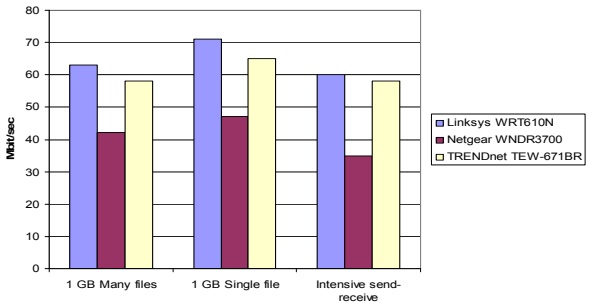


Figure 19. 2.4 GHz before processing

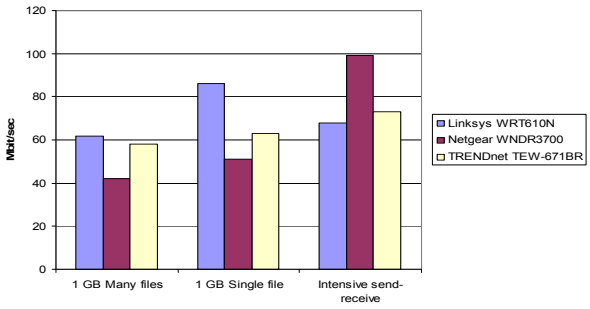


Figure 20. 5 GHz before processing

In the range 2.4GHz we received max rate in transmission regime from adapter (about 71 Mbit/s) for POP Linksys WRT610N. Receiving rate is a little smaller – on the order of 61 Mbit/s. The second indicator in the POP TRENDnet TEW-671BR. Worst performance in terms of in the POP Netgear WNDR3700. The second indicator in the POP TRENDnet TEW-671BR. Worst performance in terms of in the POP Netgear WNDR3700. In the range 5GHz, we received max rate (about 104 Mbit/s) for POP Netgear WNDR3700 only on test Intensive send-receive. When transferring files, the best result shows an POP Linksys WRT610N. After processing premises speed increased by 15-20%.

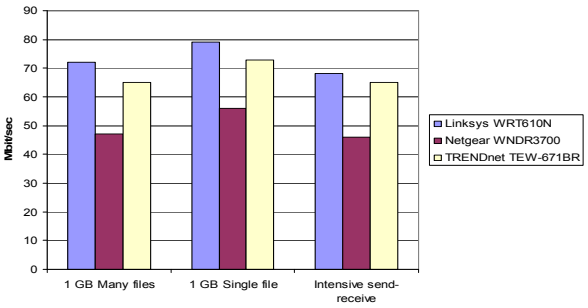


Figure 21. 2.4 GHz after processing

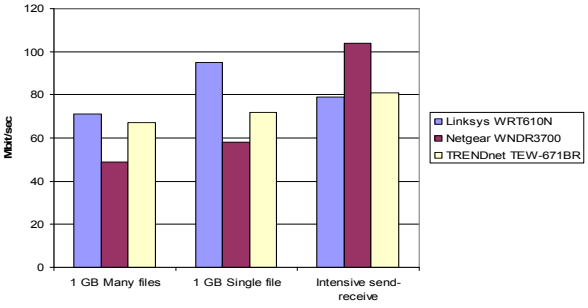


Figure 22. 5 GHz after processing

VI. CONCLUSIONS

The results of the modeling calculations carried out by the authors allow to draw a conclusion that the layered environments parameters control is possible with the helt of electromagnetic waves of vertical and horizontal polarization in the range of 2.4 GHz. To determine the dielectric layers were suggested the method of using half-wave antenna. The given theory we used for the treatment premises. The model calculations results were approved in the experiments. The experimental results at frequencies of 2.4 GHz and 5 GHz have shown that after a special treatment of premises rate increased by 5-10%.

[1] Alexey Lagunov. Increasing the Speed of a Wireless Network by Processing Indoor // Proceedings of the Seventh International Conference on Wireless and Mobile Communications (ICWMC'11) June, 2011. ThinkMind™ Digital Library. ISBN: 978-1-61208-140-3. — pp. 277-284 (http://www.thinkmind.org/index.php? view=article &articleid=icwmc_2011_13_20_20239)

[2] L.M. Brehovskih, Waves in sandwich mediums. – M.: Pub. AS USSA, 1956G.

[3] A.R. Duma, V.I. Dorohov, and A.S. Shostak, Radiowave quality monitoringof parameters of dielectric materials on the basis of measurement of an impedance of linear aerials // Flaw detection. – 1986. – N1. – pp. 54-61.

[4] E. Zajtsev, The Technique of testing of routers // http://www.ixbt.com/comm/router-mothod-2-6.shtml [retrieved: May 2012]

[5] IxChariot // http://www.ixiacom.com [retrieved: May 2012]

 ISBN: 978-1-61208-203-5